Introduction to image sequence coding

The 70-80 decades witnessed the advent of many novel telematic and television services. These services have introduced in everyone's daily routine the handling - in the communication/consultation sense - of an increasing number of pictures at various levels. Obviously, this "iconic revolution" has created between researchers, engineers and users a synergy to design, study and implement picture transmission or storage systems which make up these services. Thus a large number of studies have been launched to assess:

- **quantitatively,** the informative content (in the information theory sense [7]) of the picture signals used; in this regard, the task was essentially to determine the minimal information necessary, and strictly useful, to the complete understanding of a message being transmitted or stored.
- **qualitatively**, the visual content (in the perception theory sense [15], [2]) of these pictures; all along this book, we shall see that perception models have to be explicitly or implicitly taken into account in the very concept of the algorithmic techniques used, since in all the applications and services considered herein (see Appendix 1B), human operators (viewers, tele-observers, storage system users...) are the in-fine assessors of the quality of the service offered. To summarize this fundamental aspect, it appears evident that any transmitted or stored picture is meant to be some day viewed or consulted, *i.e.*, to be qualitatively assessed.

1 Digital picture sequences

1.1 Standard definition of a visual scene

Our visual system, or any other sensor present in a continuous natural environment (3-D + t) captures the radiosity, $F(x, y, t, \lambda)$, emitted by the sum of all light sources and reflections, where:

- $Im(F)$ = Image of (F) belongs to a continuous radiosity space. Passing from continuous representation to discrete representation of the function F (picture) constitutes the quantization phase of the image signal.
- (x, y) designate the continuous coordinates of the localization on the retinal plane (sensor = the eye) or on the sensitive surface (sensor = camera) of a local information. The (x, y) coordinate system derives, through a projection model (described in paragraph 1.1.2) from a 3-D coordinate (X, Y, Z) system. These continuous coordinates are sampled (see 1.1.4) in discrete positions $((x_s, y_s) \in \mathbb{Z}^2)$ henceforth written as $(x_s = i, y_s = j)$.
- the spectral space indexed by λ, theoretically continuous and infinite, happens to be restricted, for television applications, to the [400 nm, 700 nm] interval of the visible spectrum, and the most represented by a discrete set of three chromatic components according to the colorimetry theory [17]. For other applications (medical, satellite,

military imaging...) other spectral channels are used (ultrasound, infrared), sometimes simultaneously (multisensor imaging).

- the time axis t, sometimes processed separately from the spatial coordinate axes, is sampled at discrete instants ($t = k \in \mathbb{Z}$) by solving the heuristics that take into account:
  - the dynamic properties of our visual system (see Chapter 2),
  - the spatial resolution versus time resolution compromise.

  In parallel to this sampling of the time axis, a direction of course has to be defined for it (see Section 1.5), as well as a spatio-temporal structure.

So, by using all these discretization operations, any visual scene can be described through a spatio-temporal flow of discrete picture data called pixels or pels (picture elements), henceforth written $I(x, y; t)$ or $I(i, j; k)$ where:

- $(x, y; t) = (i, j; k) \in \mathbb{Z}^3$
- $Im(I)$ = Image of (I) $\in$ interval of $\mathbb{N}^n$, if n color components are available.

1.2 Projective system (3-D+t) $\rightarrow$ (2-D+t)

Figure 1 describes the example of the mostly used perspective system which complies with the lenticular optical model for sensor modelling.

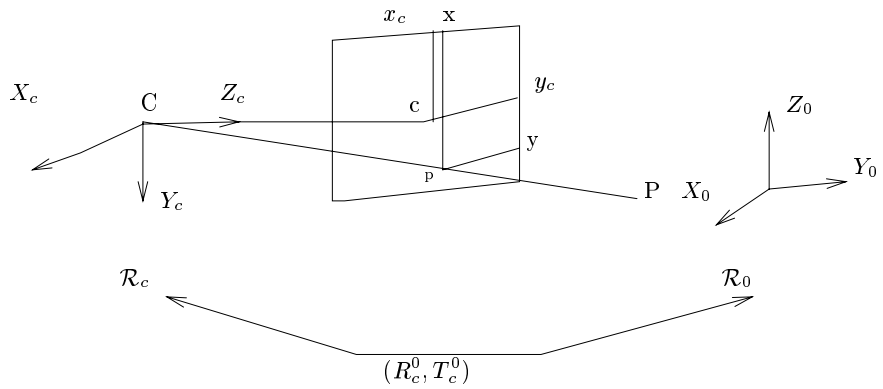


Figure 1: Projective system: example of the perspective projection

It permits changing from a system of continuous or discrete 3-D coordinates to a 2-D coordinate system (the t variable has been intentionally masked here) through the following geometrical relations:

- **Absolute coordinates**

  If the sensor is in motion, it may be useful to locate any point P of the scene in relation to an absolute reference $\mathcal{R}_o$, and not only in relation to the $\mathcal{R}_c$ reference linked to the sensor. The relations between the coordinate systems are instantaneously (at each discrete instant $t = k$)

$$\mathcal{X}_c = \mathcal{R}_c^o \mathcal{X}_o + \mathcal{T}_c^o \tag{1}$$

where $(\mathcal{R}_c^o, \mathcal{T}_c^o)$ designates the extrinsic parameters (rotation + translation) of the calibration stage, and, $\mathcal{X}_c$ and $\mathcal{X}_o$ refer respectively to the camera and scene system of coordinates.

- **Pixel coordinates**

  Once this identification $(\mathcal{R}_c^o, \mathcal{T}_c^o)$ completed, we reason henceforth in terms of coordinates linked to the $\mathcal{R}_c$ reference in pixel coordinates

$$\begin{array}{rcl} x_p & = & x_c + F_x \frac{X_c}{Z_c} + K_d F_x \frac{X_c}{Z_c} \left(\frac{X_c^2}{Z_c^2} + \frac{Y_c^2}{Z_c^2} \right) \\ y_p & = & y_c + F_y \frac{Y_c}{Z_c} + K_d F_y \frac{Y_c}{Z_c} \left(\frac{X_c^2}{Z_c^2} + \frac{Y_c^2}{Z_c^2} \right) \end{array} \tag{2}$$

  where $((x_c, y_c), F_x, F_y, K_d)$ designate the intrinsic parameters of the sensor model and represent respectively:

  - (x_c, y_c), the pixel coordinates of the projection of the optical centre C
  - F_x, F_y, ratios between the sensor focal parameter f and the basic horizontal and vertical sizes of the sampled pixel

$$F_x = f/l_x \qquad\qquad F_y = f/l_y \tag{3}$$

    with l_x, l_y, coefficients proportional to the size of the elementary pixel
  - K_d: radial distortion of the sensor.

  Usually in a number of studies, radial distortion is ignored and the F_x, F_y coefficients are normalized ($F_x = F_y = 1$), which leads to the common relationships

$$\begin{array}{rcl} x & = & x_p - x_c \\ y & = & y_p - y_c \end{array} \begin{array}{rcl} & = & X_c/Z_c \\ & = & Y_c/Z_c \end{array} \tag{4}$$

1.3 Representation of color pictures

According to the colorimetry theory [17], [19], [14], it has been shown that any color can be represented by a combination - often linear, substrative or additive - of three chromatic stimuli. The C.I.E. (*Commission Internationale de l'Eclairage*) in 1931 defined the system of chromatic coordinates based on primary colors: Red = 700 nm, Green = 546.1 nm, and Blue = 435.8 nm, whose spectral curves are shown in Figure 2. So the whole radiosity function $F(., \lambda)$ will be represented by a vector of color components such as

$$\begin{array}{rcl} \mathcal{R}(.) & = & K_R \int F(., \lambda) r(\lambda) d\lambda \\ \mathcal{G}(.) & = & K_G \int F(., \lambda) g(\lambda) d\lambda \\ \mathcal{B}(.) & = & K_B \int F(., \lambda) b(\lambda) d\lambda \end{array} \tag{5}$$

where $r(\lambda), g(\lambda), b(\lambda)$ constitute the spectral sensibility curves. Any associated reversible linear combination could be also used.

The systems of chromatic coordinates most used (see [17]), especially (Y, I, Q) or (Y, DR, DB) coordinates seek to decorrelate the Y luminance information (gray levels) from the other complementary chrominance data. For example, the (Y, I, Q) system is defined by the matrix relationship

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{6}$$

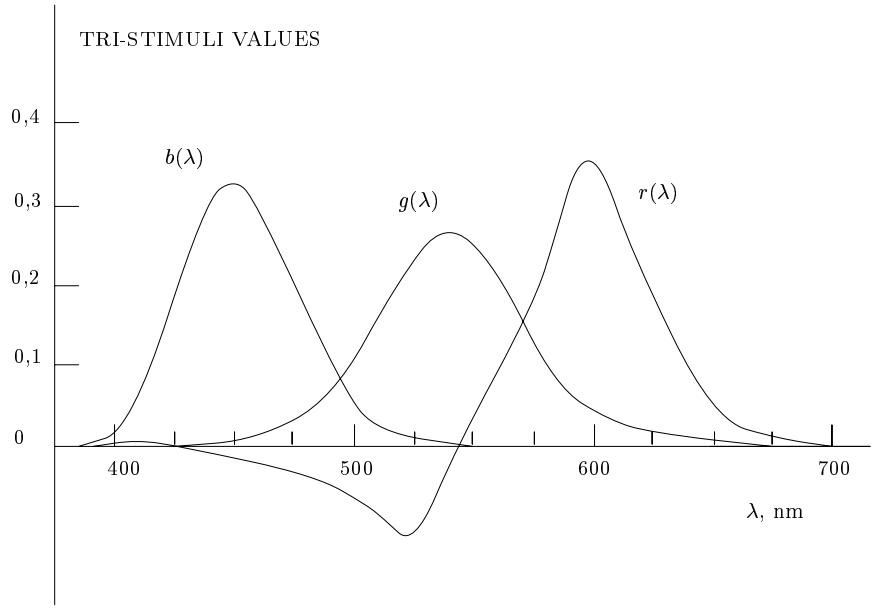


Figure 2: Color spectra

and the $(Y, DR = R - Y, DB = B - Y)$ system is defined by

$$\begin{bmatrix} Y \\ DR \\ DB \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.701 & -0.587 & -0.114 \\ -0.299 & -0.587 & 0.886 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{7}$$

The motion analysis procedures evoked in the following chapters most often use only the Y component of the initial radiosity signal.

1.4 Sampling of digital picture sequences

After the 3-D $\rightarrow$ 2-D projection, the picture coordinate system defined by Equations (2) and (4) evolves within a continuous space ($(x, y) \in \mathbb{R}^2$). The sampling process, if it follows Shannon-Nyquist's theorem, with the hypothesis of bandlimited spectrum, permits discrete localization of the picture information without data loss [5], [12]. The 1-D version of Shannon's theorem reproduced hereafter can be easily extended to any multidimensional signal sampling.

$$(x, y) \in \mathbb{R}^2 \longrightarrow (i, j) \in \mathbb{Z}^2$$

Shannon's theorem for a 1-D signal:
If a signal $I(x)$ is bandlimited to a frequency range W, that is, its spectrum $G(\omega_x)$ is equal to zero for frequencies $|\omega_x| \geq W$, then $I(x)$ can be completely specified by samples taken at the Nyquist sampling rate of $2W$ samples per unit time.

This 1-D version of Shannon's theorem can be easily extended to any multidimensional signal sampling. Appendix 1B gives various resolutions (and associated sampling frequencies) for the different levels of quality television services considered. For example, a digital TV signal sampled at 13.5 Mhz for luminance and at 6.75 Mhz for chrominance respects the

working hypothesis - which is a mere approximation - for the bounds of the luminance and chrominance spectra at 6.75 Mhz and 3.375 Mhz respectively.

If the reasoning applied to the (x, y) coordinates could in theory be extended to the spatio-temporal situation (x, y, t), it appears that the notion of infinite temporal horizon that would permit defining a spatio-temporal spectrum (2-D+t) has no real physical meaning. So sampling of the temporal axis t results more from a heuristic (as illustrated in Figure 3) which weights the notions of picture services, the spatial and temporal resolution compromise, and the visual system properties (see Chapter 2). The transverse abaci designate the isobaud curves (constant data flow) for various resolutions.

The test sequence database chosen for standard experiments in image sequence coding to illustrate this book is presented in Appendix 1C; it tries to contain various sequences representative of the various types of animated picture services mentioned hitherto.

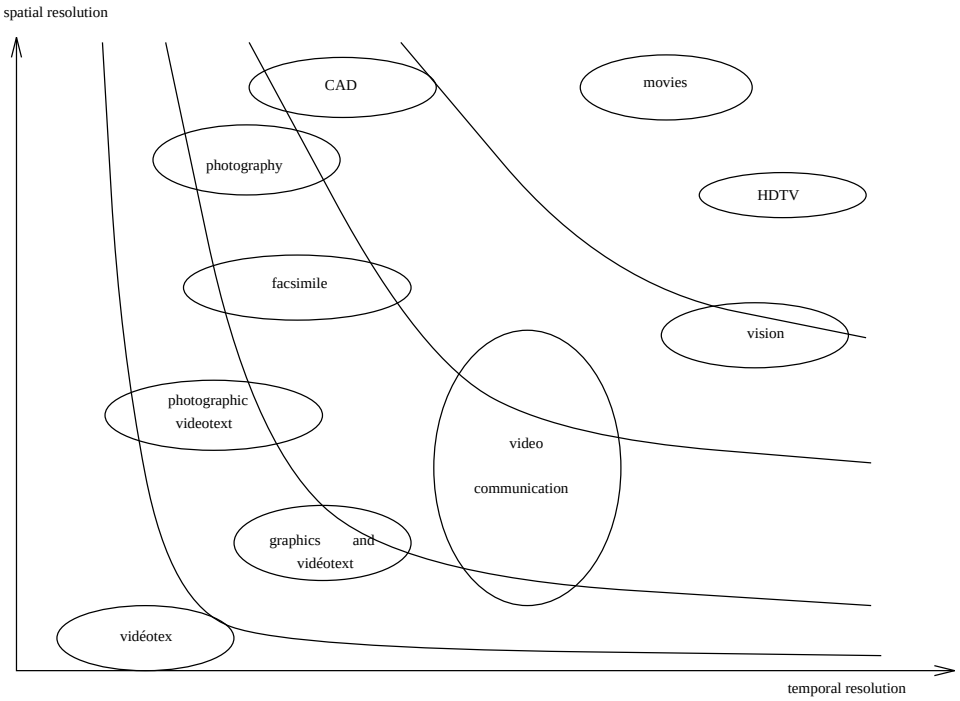


Figure 3: Spatial versus temporal resolution for different image-based services

1.5 Sampling structures for picture sequences

As a complement to the notion of sampling frequencies, spatial and temporal, it is also necessary to define that of the spatio-temporal sampling structures which will permit localizing the positions of discrete pixels in the volume of data $I(x, y; t)$.

A spatio-temporal sampling structure $\mathcal{S}_s$ will be perfectly characterized by defining the following parameter vector (see Figure 4)

$$\mathcal{S}_s = (\lambda_i, \Delta\phi_H, \Delta\phi_I, \Delta\phi_L, \Delta\phi_T)$$

where λ_i designates the interlace factor (a picture being composed of λ_i interlaced fields) and where the other structural parameters are coded horizontally according to the usual line-to-line, field-to-field and picture-to-picture scanning of picture signals.

- The usual interlace modes are:
  - $\lambda_i = 1$ (**progressive mode**) for which the field notion does not exist (nor $\Delta\phi_T$)
  - $\lambda_i = 2$, **interlaced mode** of a $2:1$ factor for which a picture sequence will be composed of two interlaced sequences of respectively even and odd fields.
- $\Delta\phi_H$ represents the horizontal offset between two successive pixels of the same line, and therefore corresponds to the sampling period resulting from Shannon's theorem (see Paragraph 1.1.4).
- $\Delta\phi_I$ represents the horizontal offset of the pixel samples of the same line between two successive frames. Comparative studies have tried to identify the respective effectiveness of fixed structures ($\Delta\phi_I = 0$) compared to mobile sampling structures ($\Delta\phi_I \neq 0$). In addition to their greater complexity, the latter more clearly display the artifacts linked to the sampling process itself (spectral foldover, aliasing) and those linked to quantization (granular noise, for instance) which become mobile (if $\Delta\phi_I \neq 0$), even in the absence of any real motion in the scene. Therefore, fixed structures are always used in general cases.
- $\Delta\phi_L$ represents the horizontal offset between the pixel samples located on two consecutive lines of the same field (if $\lambda_i \neq 1$), or the same frame (if $\lambda_i = 1$).
- $\Delta\phi_T$ represents the horizontal offset between pixel samples located on two consecutive lines of the same frame ($\Delta\phi_T = \Delta\phi_L$, if $\lambda_i = 1$).

Conventionally, and within the fixed structure family, only a few sampling structures were partially investigated:

- **the orthogonal structure** $\Delta\phi_I = \Delta\phi_L = \Delta\phi_T = 0$
- **the quincunx line structure** $\Delta\phi_L = \Delta\phi_T = \frac{\Delta\phi_H}{2}$
- **the quincunx field structure**

$$\Delta\phi_L = 0 \text{ (each field owns an orthogonal structure)}, \Delta\phi_T = \frac{\Delta\phi_H}{2}$$

The various fixed sampling structures were subjectively (psycho-visual tests) and objectively (computational complexity) compared. Briefly, what was observed can be summarized as follows:

- computational simplicity when orthogonal structures (field-wise or frame-wise) are used; the inter-line, inter-field or inter-frame interpolation methods are equally simplified. This is also true for the primitive matching methods, or pixel motion estimation methods.
- the quincunx line structure achieves the best compromise (with otherwise fixed spatial and temporal sampling frequencies) between fixed frame restitution and that of animated sequences, but it retains complexity in digital post-processing.
- the quincunx field structure tries to conciliate good temporal restitution and minimal computational complexity linked to field orthogonality.

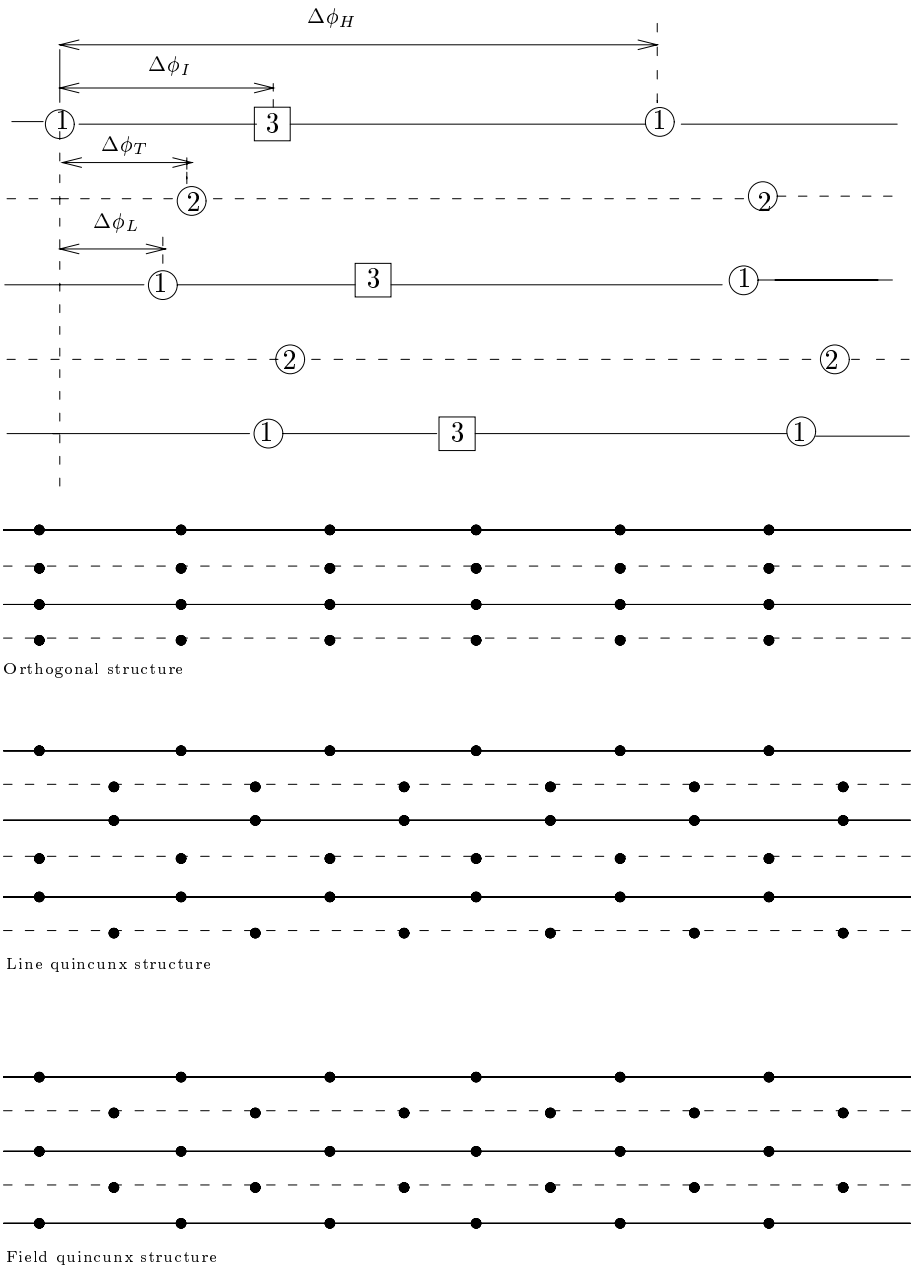


Figure 4: Sampling structures

As a conclusion to these comments on the spatio-temporal sampling of the picture sequence signal, it is worth reminding that a CCIR(*Comité Consultatif International des Radiocommunications*) recommendation was unanimously agreed upon for digital television signal specification (Recommendation CCIR-601, May 1982). This CCIR-601 standard is compatible with both existing analogic 625-line (Europe) and 525-line (USA, Japan,...) standards. It specifies the following sampling and quantization parameters:

- Coordinate system: Y, DR, DB
- Luminance sampling frequency: $f_Y = 13.5$ Mhz

- DR, DB component sampling frequency: $f_{DR} = f_{DB} = 6.75$ Mhz
- One-byte quantization: luminance range restricted to 220 values $\in [16, 235]$ for Y, and 225 values $\in [-112, 112]$ for DB and DR
- Orthogonal sampling structure
- Number of useful pixels per line: 720 pixels
- 2 : 1 interlace factor
- Temporal sampling frequency: 50 fields/sec or 60 fields/sec
- Number of lines: 625 or 525 lines/frame

This standard is identified in the hierarchy of digital signals by the standard 4:2:2 denomination (see Chapter 9). It corresponds to a 216 Mbit/s data flow.

1.6 Following the time axis

The notion of image sequence transmission in a communication system naturally induces that of a temporal causality. In this case, the time axis is followed along a single direction in increasing instants $t_0 - N, \ldots, t_0, t_0 + 1, t_0 + 2, \ldots, t_0 + N, \ldots$ from the past to the future (Figure 5). The part of image sequence already transmitted (which is the only piece of information known to the receiver of a transmission system) will be called causally reconstructed sequence.

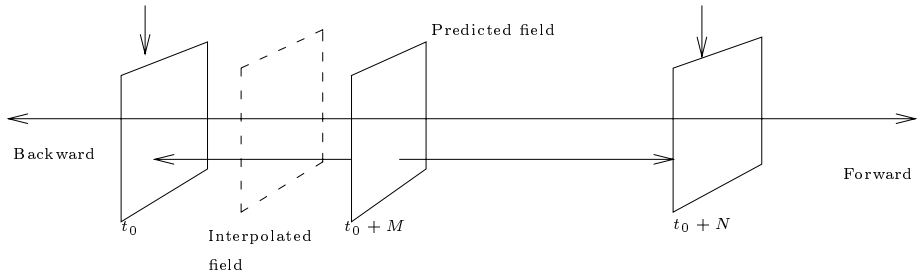


Figure 5: Temporal causality

In a storage/retrieval system, systematically reconstructing a picture located at t_0 by reconstructing the complete causal sequence is of course ruled out. To the notion of temporal causality is added that of a GOP (Group of Pictures). A $(N + 1)$ size group of pictures is located on the time axis between to reference pictures (t_0) and $(t_0 + N)$ for which instantaneous random access is possible (reconstructing images without any temporal causality constraint). Within this group of pictures, the time axis can be followed either forward or backward : such notions, necessary to retrieve moving sequences, have already been included in the MPEG standard (see Chapter 9). We shall study these notions in more detail when we introduce the motion-based interpolation techniques, or the conversion of standards using the same concept.

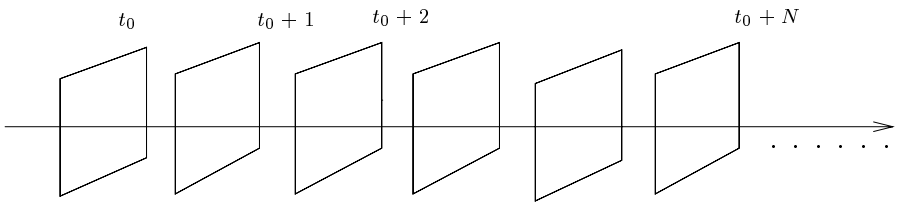


Figure 6: Temporal axis exploration direction

1.7 Intermediate conclusion

All the concepts and techniques linked to the sampling and quantization process have allowed us to represent visual data (or those perceived by a sensor) as a set of quantized discrete values $I(i, j; k)$ localized in space and time. Obviously, approximations made throughout the sequencing of these sampling techniques produce artefacts (discrete signal and visual artefacts) which cannot be remedied or reversed, and which will have to be taken into account when introducing algorithmic post-processing (*e.g.*, motion feature extraction) and during the digital signal reconstruction phases. We give hereafter a few examples of such artefacts, which occur in particular in the digital television sampling standard (625 or 525 lines/50Hz/2:1, $F_{sY} = 13.5$ MHz).

- **Spatial spectrum overlap**: the bandlimited spectrum constraint of Shannon's sampling theory is never achieved, indeed, with real signals. Imperfect filtering of residual spatial high frequencies may produce such artefacts.
- **Temporal spectrum overlap**: this phenomenon is specially visible on structures with periodical, wide-amplitude motions. The usual case of such a defect is illustrated by the temporal stroboscopic effect linked to the discrete temporal sampling (50 frame/s, or even 24 picture/s for cinema sequences) that are easily perceived: *e.g.*, lenticular wheel of a bicycle appearing to turn in reverse of the cyclist's motion.
- **Flicker effect**: this defect is to be related to a too small value of the field frequency in relation to the temporal characteristics of the visual system (Chapter 2). The flickering phenomenon is only critical during the restitution of moving sequences. Some systems have evoked the possibility to introduce distinct temporal sampling frequencies:
  - at the level of acquisition and analysis: dependency on the dynamic characteristics of the scene to be perceived
  - at the restitution level: dependency on the visual system characteristics
- **Interline scintillation**: the interlace process is responsible for this defect which can be seen essentially around the horizontal or slightly oblique contours.
- **Line visibility**: the low vertical resolution in terms of number of lines reduces the fine perception of details. Let us remember that the vertical resolution is reduced, compared to its horizontal counterpart, by applying Kell's factor (4/3 or 16/9 according to standards or services).

All these defects are countered either by specific processing, or by defining improved picture services (DATV, HDTV, ...).

2 Transmission and storage systems

For all the applications and services mentioned in Appendix 1B, a communication and/or storage system can conventionally be defined as the assembly of the modules in Figure 7.

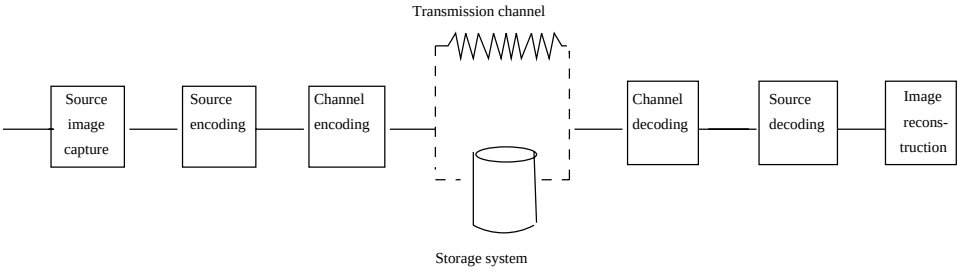


Figure 7: A communication system

2.1 Coder-transmitter part

2.1.1 A source-coding module

This module, described in more detail in Figure 8, forms the main part of a compression system [10], [17]. In that sense, the algorithms made up in picture coding and those described in this book for motion analysis in picture sequences, are the design basis for such a source coding module. As a rule, this module is decomposed into two phases: analysis and extraction of the data necessary for transmission or storage.

The analysis phase can use any algorithmic technique developed for digital picture processing [19], [11], sometimes outside of the strict Picture Coding context. We thus find a possible structuring of the methodological tools used for:

primitives: pixel, contour, region, feature grouping, object

type of processing: transformation, filtering, and for motion analysis: detection, estimation, segmentation, labelling (classification)...

resolution methods: statistic, geometric, syntaxic approaches...

A number of general knowledge books have been dedicated to the description of these tools and can therefore provide this analysis stage with efficient algorithms. At the end of this first phase, no data compression has yet been achieved, and more critically yet, this analysis phase contributes to the information knowledge that is available on the perceived picture scene, thus creating additional information in addition to initial picture data.

To summarize, this analysis phase generates, for any pixel spatio-temporally localized at (x, y, t), a pair of data ($I(x, y; t)$ and $C(x, y; t)$) where:

- $I(x, y; t)$ is the original picture information

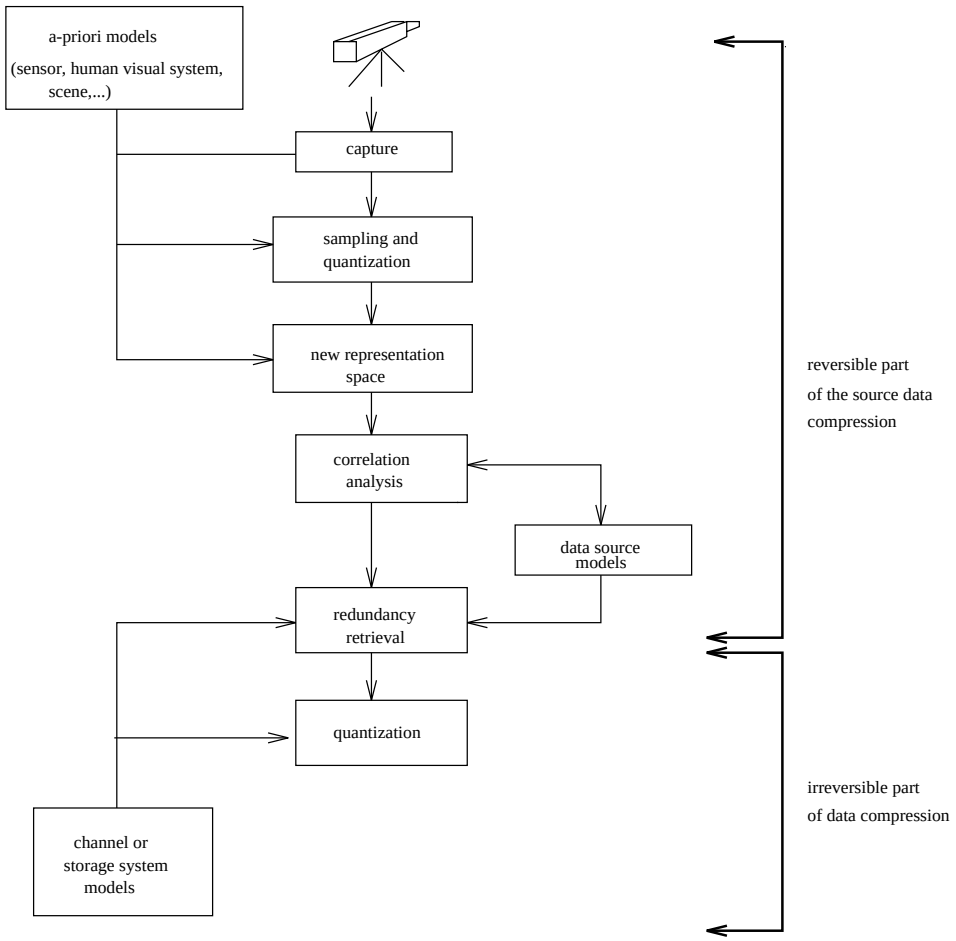


Figure 8: General scheme of a data compression system

- $C(x, y; t)$ is the characterization of the original picture by one of the previously mentioned processings.

The second phase uses these information pairs to extract the useful information - called *spatio-temporal innovation* - for accurate restitution (quality criterion to be defined) of reconstructed picture signal. For example, in differential coding diagrammes (see Typology of compression methods in paragraph 1.1.5), we have the following relation

$$i(x, y; t) = I(x, y; t) - f_p(Mem(I), C(x, y; t)) \tag{8}$$

where:

- $I(x, y; t)$ = image signal
- $i(x, y; t)$ = spatio-temporal innovation
- $f_p(., .)$ = predictive function
- $Mem(I)$ = causally reconstructed and stored image signal
- $C(x, y; t)$ = extracted spatio-temporal characterization.

The data compression achieved at this level consists of a **reversible** coding (named lossless techniques) by analysing the correlations present in the signal, and extracting redundancies.

The source-coding phase is then completed by (scalar or vector) quantization of the innovation signal $i(x, y; t)$ [6], and sometimes by quantization of the characteristic information $C(x, y; t)$, if it is to be transmitted. This processing achieves a decisive step of **irreversible** coding: some objective information is definitively lost at that level. All is done to reduce or prevent subjective information, in the visual or semantic sense.

Finally, it is to be noted that this source-coding module, especially in terms of bit rate, will result from two factors:

- the effectiveness of $C(x, y; t)$ to compress as much as possible the picture-signal information $I(x, y; t)$;
- the "compact" representation of $C(x, y; t)$, if it is transmitted. Two examples are referred to in this book: 1) necessity to obtain a compact representation of the region boundaries resulting from the spatio-temporal segmentation phase, 2) tentative compact representation of a field of motion vectors resulting from a motion estimation phase.

2.1.2 A channel-coding module

Studying algorithmic techniques in Motion Analysis - the subject of this book - interferes but little with the techniques involved in Channel Coding. We shall therefore not go into any detailed description of these techniques, which are identical to the channel coding techniques described in general studies on fixed image coding.

The channel-coding module produces a binary data stream protected by error detecting/correcting codes [18], [3], [1]. The techniques involved are based on the information theory - statistical coding (VLC), channel capacity and coding- and on binary algebra (error-correcting codes, cryptographics, ...). The contribution to this field, specific to picture sequence coding, lies in the study of $C(x, y; t)$ statistics, considered as motion vectors, and the tentative specification of special error-protecting codes in relation to this particular information.

2.2 Decoder-receiver part

Three processing modules, dual to those previously described for the transmitter, exist at the reception level of a communication system.

1. The channel decoder module extracts, even corrects, if an error was detected, the useful information from the complete message transmitted.
2. The source decoder module reconstructs the complete signal $I(x, y; t)$ by adding to the transmitted spatio-temporal innovation the redundant portion of the signal that was not transmitted. This recomposition phase is classically called Reconstruction or Synthesis of the picture sequence, by analogy with the analysis module present at the source encoding level. A communication system can then be described by the Analysis/Synthesis chain within which non-linear and irreversible quantization processings are interpositioned.

3. The reconstructed sequence restitution module must follow a protocol linked to the picture service targetted and to the associated restitution quality, thus defining the resolution spatial and temporal frequency characteristics, the objective or subjective evaluation criteria (Appendix 1A).

3 Extraction and exploitation of inter-frame correlations

Several processing phases try to exploit inter-frame correlations for coding purposes. We would essentially distinguish motion detection from motion estimation and compensation.

3.1 Statistical properties of the picture sequence signal

It has now been accepted that the picture sequence signal is non-stationary both in space and in time. Thus, for a given picture (see Table 1), the luminance probability density function is not uniform; it varies with the picture, and even temporally within a sequence. The experiment in Table 1 is carried out with *Split Screen* sequence. Y, A and C pixel localizations denote, relative to the current pixel X, respectively the same pixel at the previous frame, the previous and the upper pixels in the same frame. The entropy of this distribution $\{p_i\}$ for $i \in [0, 255]$, the luminance gray scale, is expressed as

$$H(p) = -\sum_{i=0}^{255} p_i \log_2 p_i \tag{9}$$

In the sense of the information theory, it quantifies the minimal length of a binary symbol codeword that permits representing the achievements of this probability law; we observe that the codeword used for uniform quantization is never very far from 8 bits (in this case).

Frame Entropy	21	22	23	24	25	26	27	28	29	30
H(X)	6.08	6.08	6.08	6.08	6.09	6.12	6.12	6.09	6.06	6.05
H(X,Y)/2	4.58	4.61	4.53	4.57	4.69	4.79	4.80	4.77	4.76	4.72
H(X/Y)	3.18	3.23	3.05	3.16	3.42	3.64	3.62	3.56	3.57	3.51
H(X−Y)	3.89	4.01	3.69	3.88	4.26	4.63	4.61	4.55	4.59	4.46
H(X,A)/2	4.91	4.92	4.93	4.92	4.92	4.92	4.94	4.93	4.89	4.92
H(X/A)	3.89	3.85	3.86	3.85	3.84	3.81	3.84	3.85	3.80	3.87
H(X−A)	4.52	4.55	4.58	4.55	4.51	4.47	4.50	4.51	4.41	4.52
H(X,C)/2	4.76	4.77	4.77	4.76	4.76	4.77	4.80	4.79	4.79	4.80
H(X/C)	3.59	3.60	3.61	3.60	3.57	3.57	3.62	3.64	3.65	3.69
H(X−C)	4.37	4.39	4.41	4.38	4.31	4.28	4.35	4.39	4.37	4.44

Table 1: Temporal stability of intra/interframe statistics by evaluating zero-order, joint and conditional entropies of the image signal.

In contrast, block statistics and extending the entropy computations to the case of joint $(H(X,Y)/2, H(X,A)/2, \ldots)$ and conditional $(H(X|Y),\ H(X|A),\ \ldots)$ probability functions (see Table 2), clearly show the existence of local dependencies (or redundancies) whose exploitation will permit the signal source compression (suppression of redundancies and transmission of the innovation).

Sequence / Entropy	Mobile	Renata	Car	Rubix	Miss	Sphere	Split	Kiel
H(X)	7.62	7.42	7.34	4.11	5.77	4.35	6.09	7.22
H(X,Y)/2	6.67	6.47	6.76	3.19	3.95	3.12	4.69	6.69
H(X/Y)	5.85	5.60	6.33	2.36	2.21	1.99	3.42	6.28
H(X−Y)	6.54	6.29	7.09	3.58	2.89	2.88	4.26	7.26
H(X,A)/2	6.49	6.50	6.45	3.09	4.71	3.18	4.92	6.38
H(X/A)	5.48	5.67	5.67	2.13	3.72	2.05	3.84	5.58
H(X−A)	6.06	6.41	6.19	3.01	4.12	2.71	4.51	6.33
H(X,C)/2	6.66	6.67	6.65	3.29	4.71	3.30	4.76	6.62
H(X/C)	5.84	6.04	6.12	2.57	3.77	2.29	3.57	6.16
H(X−C)	6.52	6.89	6.83	3.60	4.20	3.27	4.31	7.14

Table 2: Intra/interframe statistics by evaluating zero-order, joint and conditional entropies of the image signal for different image sequences

Tables 1 and 2 illustrate, from several examples of picture sequences with different characteristics in terms of spatial and motion contents, the statistical properties of the sequence signal. A few general comments can be made in view of these experimental results. Joint or conditional entropies are smaller than zero-order entropies of the original signal which appears quite stable temporally. Temporal redundancies are greater than spatial ones, both in horizontal and vertical directions. This difference is increased when scenes with low magnitude motions are tested. Only the so-called *Kiel* sequence shows the contrary and, as far as a global (camera) motion is concerned, it creates globally uncorrelated data in the temporal dimension. With regard to temporal statistics, entropic computation of $H(X,Y)/2$ or $H(X|Y)$ performed on two pixel locations along the same temporal axis makes sense only if they are not moving. Clearly, the temporal correlations of two points of an object moving between instants $t-1$ and t are not to be sought along a parallel to the temporal axis, but in the exact direction of pixel movement. This supposes that this motion is known. The following paragraph defines the basic concepts for any motion analysis used for inter-frame coding, with **motion compensation**.

3.2 Motion analysis and coding

With a view to extract as many inter-frame redundancies as possible, it is necessary to analyse the dynamic content, *i.e.*, the motion present in the picture [8], [22]. This analysis can be performed at various levels of knowledge and complexity, as shown in Figure 9. Briefly, we shall only evoke the analysis of motion along a short temporal horizon, restricted to two successive frames. We shall seek motion analysis phase, taking into account notions of temporal linking in a longer sequence of consecutive frames, in Chapters 3 and 7.

Motion detection phase: This process aims to extract from a sequence the moving areas and the non-moving areas, in a series of observations essentially based on inter-frame differences. The product of this process is a binary diagnosis of presence/absence of motion and a spatial mapping of mobile object masks. A pair of images can appear as insufficient to suppress detection ambiguities (overlapping, uncovering, internal area of the moving object...)

Motion estimation phase: The aim is to quantify the perceived motion information (ap-

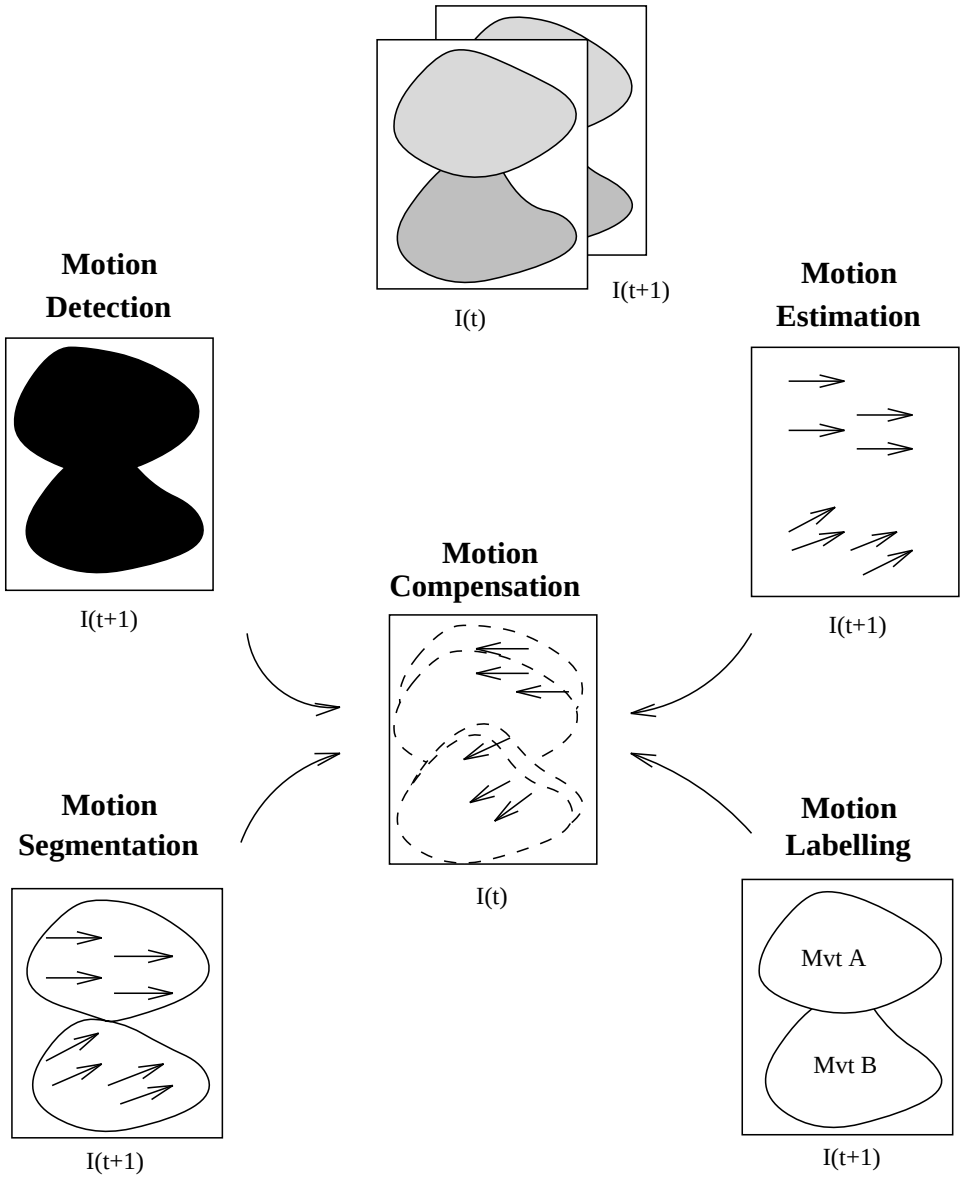


Figure 9: Motion analysis tasks for motion-compensated coding

parent velocity) through picture observation. This study phase has produced many algorithms (see Chapter 3) to measure the fields of motion vectors [9], [21]. These measurement fields are characterized by

- their density: dense field, if there is a measure applied to each pixel, sparse field, if this estimation phase is only done on a feature basis.
- the type of primitives they involve: local pixel-based motions, contour motions (1-D case), region motions (2-D case), object motion (3-D case), ...

After this identification (which implies defining models of such motions), quantitative measures are provided with a final accuracy level which can be as high as sub-pixel. In this case the motion vectors will be estimated, as an example, as 1/8th, 1/16th pixel/frame.

Motion segmentation phase: As for the detection phase, it is important to analyse a scene in terms of homogeneous regions, as a motion criterion. Different from motion detection, a motion-based segmentation algorithm ensures the segmentation of the scene into regions with distinct and identifiable movements.

The "Chicken and Egg" problem: Let us remind the difficult problem of simultaneous optimization of the estimation and segmentation phases. This is a general problem, applying to spatial segmentation (based on texture attributes, for instance) as well as to the present motion-based segmentation. The dilemma is to be summarized as follows: to achieve a good segmentation into homogeneous regions in term of motion, motion parameters must have been well assessed first; but estimation algorithms are only truly effective in areas without sharp model discontinuity (in this case, motion models), which implies prior effective segmentation into homogeneous areas without model discontinuity. This naturally leads to iterative or mixed algorithms producing or taking into account these two processing notions simultaneously.

Motion interpretation phase: For many scene analysis applications, and also for those of semantic coding with preclassified motion models, it may be interesting, in terms of comprehension, hence of semantic compression, to analyse a picture sequence only in terms of motion qualitative definition: object 1 rotating, object 2 zooming, ... These high-level interpretation phases derive from the previous low-level processing phases of detection/segmentation. This type of processing is essential when, in a communication system, it is necessary to perform compression and diagnostic with the same algorithmic processing. This is the case in active monitoring systems using video communication networks.

Motion compensation phase: This processing phase is specific of the data compression schemes. Effectively, we noted in paragraph 1.4.1 that the inter-frame correlations observed on axes parallel to the temporal axis had but little significance for moving objects. **Motion compensation** of a picture implies prior analysis of motion, as previously described, and exploitation of the data so obtained to create a virtual picture $I_{comp}(x, y; t)$, which in each point, block, or region, according to the estimated motion information density, seeks in picture I_{t-1} the "displaced" radiosity information

$$I(x - D_x, y - D_y; t - 1)$$

if (D_x, D_y) designates a motion vector estimated between instants (t) and $(t-1)$ for pixel (x, y) with reference to picture t. Motion compensation thus permits using the inter-frame correlations according to the estimated motion.

4 Types of compression methods

The research topic of digital picture data compression is not new, [16], [10] and several authors [17], [13], [20] have proposed a classification of the various algorithmic methods involved. We propose hereafter a possible typology of these methods (Figure 10) by including the novel research avenues recently opened, in particular concerning the subband coding [23] and object-oriented coding.

This classification hierarchy can be interpreted as follows:

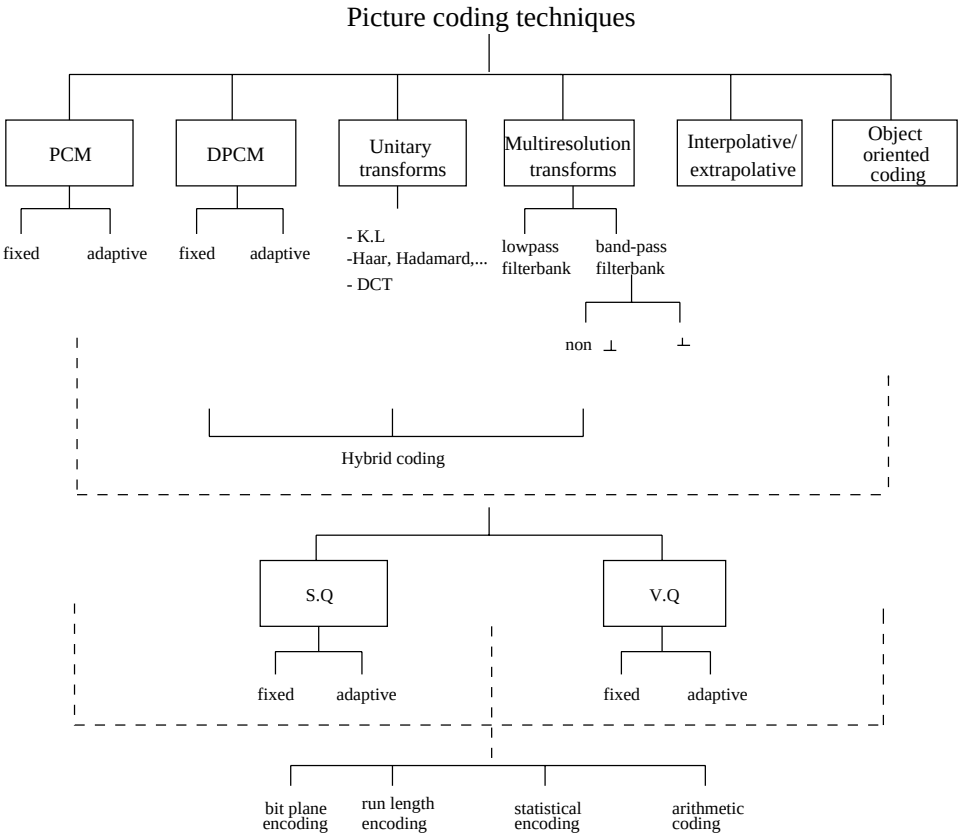


Figure 10: General classification of compression methods

- The initial step consists in defining, from the initial space of picture data, a new representation space, either predictive, or frequential, [4], [23], or semantic, permitting isolation - hence extraction - of the resident redundancies within the signal to be processed. Thus will be distinguished:

  1. **Spatial redundancies**, and in moving picture sequences, spatio-temporal redundancies, resulting from inter- and intra-frame correlations between adjacent pixels.
  2. **Frequential redundancies** - or spectral in multispectral signals - present in several correlated frequency bands.
  3. **"Object" representations** which, through high level semantic models, try to characterize local pixel data by global attributes which discriminate between object entities (contour, region, texture, motion) present in the scene.

  Historically, these various methods were not introduced simultaneously. It is now quit apparent that **hybrid schemes**, *i.e.*, algorithms which utilize sequentially several representation spaces, provide higher performance in terms of spatio-temporal and frequential redundancy extraction. Hybrid algorithms, mixing frequential representation with predictive error data from a DPCM predictive diagram, are routinely used. This has also been the case more recently for predictive multi-band algorithms.

- The second, optional step, which when used creates the **irreversibility** of the coding

process, involves the source quantization [6] of residual data, *i.e.*, the "innovation" part of the signal, produced by the preceding phase. Conventionally, this information is quantized in a scalar (SQ) or vector (VQ) manner, which implies in the latter case the existence of residual correlations between the components of the defined vector. A number of methods have tried to specify the characteristics of the quantizers:

- based on statistical models of input data; these models are assumed to be known *a priori* by prior off-line learning, or they can be adapted.
- taking into account the psychovisual properties, the visual system being the rater of the levels of signal reconstruction defined by the quantizer; this type of visually based quantization may profit greatly from an adapted frequential decomposition: visual sub-band coding or DCT transformation.

Let us remember that this quantization module is irreversible and very often not uniform. It thus creates a coding diagram with losses (lossy techniques) as opposed to the lossless techniques which do not use such a module.

- Lastly, the third and last step performs a "binary" coding (or symbolic, if the alphabet is wider), optimal in the sense of information theory. For a volume of data produced by one or both preceding phases, it is necessary to select the method best adapted to the statistics or distribution geometry of the discrete elements to be coded.

  These are:

  1. bit plane encoding
  2. run length encoding
  3. entropic coding, function of an empiric, monovariable probability density
  4. arithmetic coding, or contextual entropic coding, taking account of context conditional probability density function (*e.g.*, m-order Markovian source).

5 Coding and motion

This book being dedicated to the motion analysis techniques which aim at compressing sequences of moving frames, we would like to explain hereafter briefly at which level the notions of temporal dimension and motion are naturally introduced in the typology of compression methods previously described (see Figure 10).

5.1 Coding and temporal non-stationarity

A natural sequence of moving frames (film, television broadcast,...) is by definition non-stationary in time. A "cut" between two sequences with different filming positions and instants is an extreme example, although a frequent occurrence. Within the same sequence, the moving object appearance-disappearance phenomenon also creates temporal non-stationarities. This argues for adaptive methods with regard to:

- the choice of predictors (DPCM coding)
- the choice of quantizers
- the taking into account of statistics (non-stationary) for entropy coding.

While the picture areas with sharp spatial discontinuities (contours) are those where inter-frame coding schemes require very strong adaption and nevertheless remain critical areas, sources of reconstruction errors, likewise the picture areas with sharp temporal discontinuities (motion or temporal "changes") are those which use the motion analysis algorithms, which are the only ones capable of promptly taking into account (for instance by instantaneous adaptation of the predictive model) of a temporal non-stationarity.

5.2 Coding and temporal separability

The tridimensional space (2-D+t) of picture pixel information $I(x, y; t)$ is enriched with the third dimension (time axis), compared with the 2-D space of a static picture. This new coordinate axis can be processed:

- **in a separable manner**: we have mentioned in paragraphs 1.1.3 and 1.1.4 the specificity of the time axis with regard to spectrum and sampling; a separable processing is thus often used, for the spatial axes (x, y) and the time axis (example: 2-D orthogonal transformation and temporal prediction).
- **in a non-separable manner**: this is the case when the spatio-temporal "voxels" are directly exploited in the 2-D+t space. This occurs when a neighborhood system is spatio-temporally defined to seek the optimal predictor; likewise, some research (subband coding) use tridimensional frequential transformations in Analysis/Synthesis filter banks defined in the corresponding 3-D spectral space; these filters are often separable, which brings us back to the preceding situation.

5.3 Coding and the nature of extracted motion information

Section 1.3 defined various possible steps - of increasing complexity - of motion information extraction. To these various steps correspond several compression uses, which are summarized in the adaptive DPCM predictive techniques with the following relations:

Motion analysis	Motion-oriented coding
Change detection	Conditional replesnishment
Dense displacement estimation	Motion compensation
Estimation-Segmentation	Spatio-temporal object-oriented coding
Interpretation	Semantic coding

These various processing phases can be used sequentially.

Conditional replenishment (see Chapters 3 and 4) The motion detector producing a binary location of the mobile areas in the scene, only those areas will be refreshed in the picture memory containing the previous reconstructed picture. Temporal refreshment is therefore conditioned by the motion detection diagnosis. The innovation signal of any pixel then is the *frame difference* ,

$$FD(x, y; t) = I(x, y; t) - \tilde{I}(x, y; t-1) \tag{10}$$

where $\tilde{I}$ designates the previous picture known to both coder and decoder.

Motion compensation (see also Chapters 3 and 4) After prior estimation of a dense field of apparent motion vectors (at each point $(x, y; t)$, a $\vec{D} = (D_x, D_y)$ vector is estimated),

motion compensation creates at any point a virtual picture which corresponds to the displaced intensities of estimation motions and referenced in the $t-1$ picture. The considered innovation signal is then the *displaced frame difference*,

$$DFD(x,y;t) = I(x,y;t) - \tilde{I}(x - D_x, y - D_y; t-1) \tag{11}$$

Spatio-temporal object-oriented coding (see Chapters 3 and 8) The dynamic scene is described here as a homogeneous segmentation in the sense of a motion criterion (to which a surface model and radiosity criterion can be added). The innovation signal is composed of the dynamic evolution of these regions, taken in the wider sense, *i.e.*, including the notions of appearance/disappearance, deformation, and merging of objects.

The data manipulated involve:

- the coding of the dynamic structures linked to the segmented regions,
- the coding of identified parameters of motion (and/or surface) model,
- the coding of radiosity textures.

Semantic coding This processing level requires geometric and radiosity very high level modelling for the objects and scenes manipulated. Likewise, the Interpretation phase resulting from Motion Analysis generates a very compact qualitative description - *e.g.*, a given sequence from "gesture" language. Obviously, this can only be envisaged in restricted applications where wide *a priori* knowledge can be injected or identified by prior learning. However, there is a natural continuum between these methods and the lower level versions of the previous techniques.

5.4 Miscellaneous

To complete the description of the possible levels of introduction of "time", and not only motion, in the typology of compression methods, the following should also be mentioned:

- adaption of vector quantization [6] schemes to the temporal non-stationary situation (codebook refreshment),
- designing of interpolation/extrapolation methods in the direction of motion, with the underlying hypothesis on the motion trajectory model, if a temporal horizon of several frames is effectively used. Also in this case, the time axis (and therefore the motions related to each of them) can be followed forward and backward (see paragraph 1.1.6).

6 Plan of the book

We just describe briefly the contents of the following Chapters.

Chapter 2 will introduce some notions of perception of time-varying images and how the experimental psychovisual laws could be useful to adapt encoding schemes to the human visual system. Time-varying imagery perception involves spatio-temporal masking effects and needs to take into account the specific characterics of human vision. Psychovisual modelling tools such as receptive fields tuned to a spatio-temporal frequency band is more especially described; it appears very close to the subband formalism also introduced in image sequence coding.

Chapter 3 presents the major algorithmic tools for motion analysis following the methodological classification we have introduced in Chapter 1. We firstly focus on the difference between 2-D apparent velocities and real motions. The two-dimensional motion analysis is then described in the case of a curvilinear (moving contours) or a two-dimensional (moving regions) domain. Change detection, velocity field estimation and motion-based segmentation are sequentially addressed.

Chapters 4 and 5 will describe encoding schemes respectively using predictive motion compensating techniques and hybrid techniques also using orthogonal transformations. The aim of motion compensated prediction loop is to decorrelate the image sequence signal in the direction of (locally estimated) motion and extract the spatio-temporal innovation, which is then, as far as an hybrid scheme is concerned (Chapter 5), transformed, and in any case quantized. Chapter 4 describes in a detailled fashion the three main components of any predictive coder: predictor, quantizer and encoder. Chapter 5 presents how to optimally combine two decorrelation stages, *i.e.*, an orthogonal transformation and a motion compensating prediction. These chapters also present the rate/distortion functions obtained for different accuracies of the motion analysis stage.

Subband and multiresolution approaches are presented in **Chapter 6**. The main goal is to describe motion analysis algorithms which exploit several resolution or frequency levels of image data. These algorithms are called multigrid, multiscale or multiconstraint approaches relative to how the multiresolution analysis is carried out with motion analysis. Therefore wavelet or orthogonal subband decompositions are evoked and illustrated by hybrid subband coding.

In **Chapter 7** motion-based interpolation algorithms are shown; several applications are concerned with temporal interpolation. De-interlacing process or standard conversion need efficient tools based on motion-oriented interpolators. Two main cases will be explored according to the assumption that motion informations are transmitted or not. The MPEG standard family (see also Chapter 9) already uses these interpolators.

Object-oriented encoding schemes and especially high-level motion modelling are depicted in **Chapter 8**. We naturally extend two-dimensional and low-level motion analysis trying to estimate 3-D motion and structure information maps. Monocular and binocular (stereoscopic case) are distinguished and briefly an application to 3DTV image sequence encoding is presented.

This book concludes with **Chapter 9** where some standards for image sequence compression already established are described. It evokes three existing standards concerning the coding of image sequences; we restrict our discussion to the contribution-quality digital TV transmission (CCIR Rec. 723), storage of animated imagery on digital support (MPEG) and video-phone/-conference systems (CCITT H.261) and their extension to multimedia applications.

The authors would like to expect that all the methodological tools presented in this book (Chapters 1 to 8) would anticipate and possibly help the design of new future standards for advanced image communication systems.

7 Appendices

7.1 Appendix 1A: "Objective" quality criteria

The quality evaluation between an original image and a recontructed one after coding-decoding is surely one of the most crucial problems. Many subjective tests involving human observers are performed to evaluate these quality levels. Only "objective" quality measures are listed hereafter and obviously these criteria are absolutely not well fitted to human visual system properties. An alternative way is to weight these objective measures by some perceptual masking functions relative to a Human Visual System modelling.

If $I(i,j)$ and $\tilde{I}(i,j)$ denote respectively the original and reconstructed images then usually the following objective criteria are defined:

- **Mean Square Error: MSE**

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(I(i,j) - \tilde{I}(i,j)\right)^2 \tag{12}$$

  with M and N the horizontal and vertical image sizes.

- **Signal to Noise Ratio: SNR**

$$SNR_{db} = 10 \ \log_{10} \ \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} I^2(i,j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} \left(I(i,j) - \tilde{I}(i,j)\right)^2} \tag{13}$$

- **Peak Signal to Noise Ratio: PSNR**

$$PSNR_{db} = 10 \ \log_{10} \ \frac{255^2 MN}{\sum_{i=1}^{M} \sum_{j=1}^{N} \left(I(i,j) - \tilde{I}(i,j)\right)^2} \tag{14}$$

  if the original image signal is quantized over a 256-gray level scale.

7.2 Appendix 1B: Video formats and transmission channels

- **Video formats and applications**
  - **CIF** (Common Intermediate Format)
    - 288 active lines $\times$ 352 pixels per line (and not 360 in order to work on macroblocks of size 16 $\times$ 16)
    - Field repetition frequency: 30Hz (20Hz is also sometimes evoked)
    - The chrominances are downsampled by 2 in both directions
    - Remark: QCIF format is a restrictive format obtained by subsampling CIF video format by 2 in both directions
    - Applications: Video-phone, video-conferencing services
  - **4:2:2 Format** (CCIR-601 recommendation) (described here for the 625-line system)
    - 576 active lines $\times$ 720 pixels per line
    - Field repetition frequency: 50Hz

  - Interlaced factor: 2:1
  - The chrominances are downsampled by 2 in the horizontal direction
  - Aspect ratio: $\frac{4}{3}$
  - Applications: Digital TV

- **High-definition formats**

  Eureka-95 HD-MAC project introduces the following digital hierarchy

  - Progressive formats
    - 1250 lines × 1920 pixels per line
    - Field repetition frequency: 50Hz
    - Interlaced factor: 1:1
    - Orthogonal sampling structure
    - Subsampling factor of 2 for chrominances
    - Remarks: some others intermediate formats for the luminance signal have been introduced: "Extended Definition progressive format EDP" (960 × 625), "High-definition quincunx structure HDQ" (960 × 1250), "Extended definition quincunx structure " (480 × 625)
    - Applications: High Definition Digital TV
  - Interlaced formats: Idem as previous HD (or ED) formats with an orthogonal sampling structure and with an interlace factor of 2 for the luminance signal.

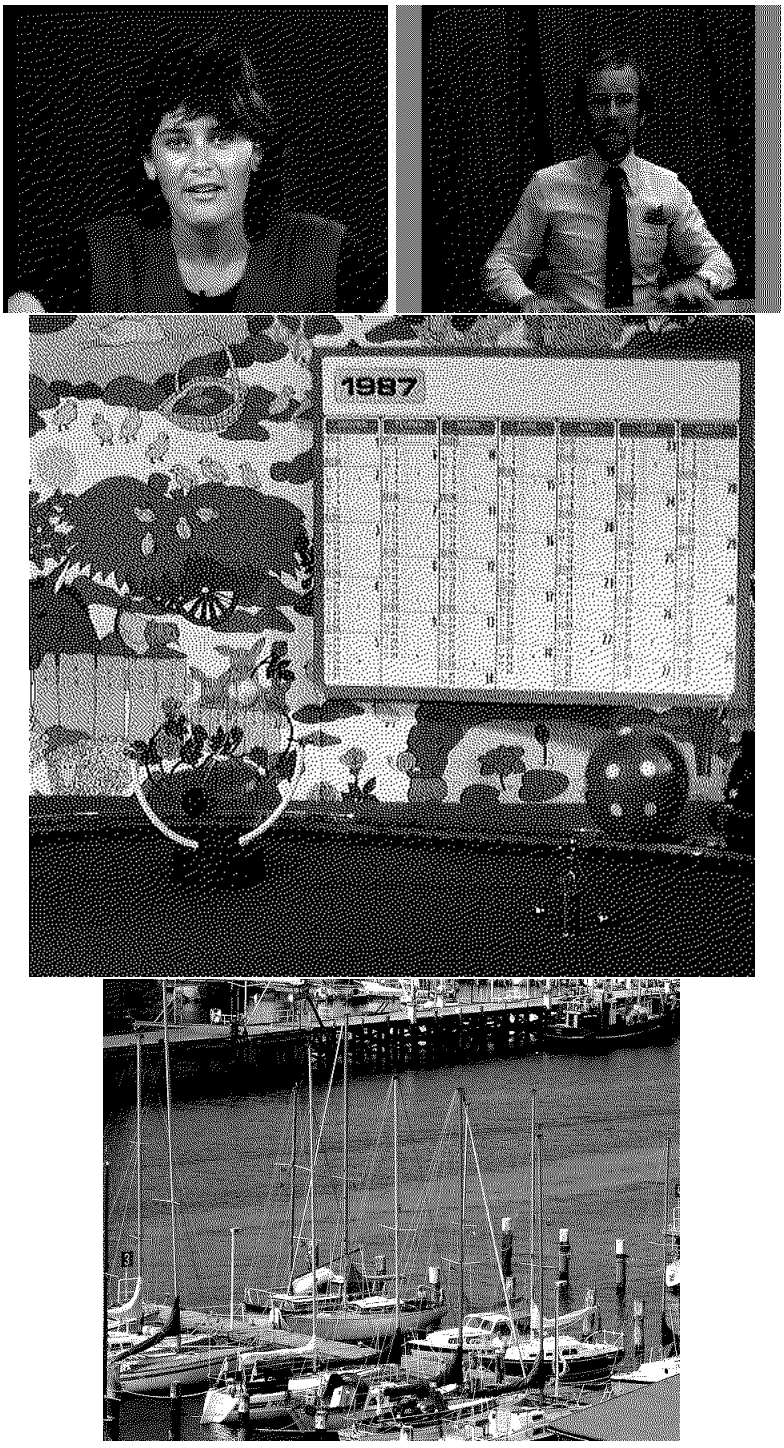
  Some more recent studies try to optimize the crucial HDTV format following others rules for HDTV sources: from 720 lines × 1280 pixels per line (at 30Hz image frequency or 60Hz field frequency) to 1080 lines × 1920 pixels.

- **Transmission channels: CCITT Hierarchy definition**

Level	Bit rate	Application
0	64 kbit/s	Videotelephony
	p*64 kbit/s	LQ Videoconferencing
1	2 048 Mbit/s	HQ Videoconferencing LQ Digital TV
2	8 448 Mbit/s	MQ Digital TV Video Recorder
3	34 368 Mbit/s	TV contribution Digital TV distribution (1 or 2 channels)
4	139 264 Mbit/s	HDTV contribution Digital HDTV distribution (1 or 2 channels)

7.3 Appendix 1C: Some examples from a test sequence database

Examples of different images corresponding to various image communication applications:
a)*Miss America* video-phone sequence: 360 pixels × 288 lines per frame, 15Hz
b)*Split Screen* video-conferencing sequence: 360 pixels × 288 lines per frame, 10Hz
c)*Mobile and Calendar* digital TV sequence: 4:2:2 format: 720 pixels × 288 lines per field, 50Hz
d)*Kiel* HDTV progressive sequence: 720 pixels × 576 lines per frame, 50Hz



References

[1] J. Anderson and S. Mohan, *Source and channel coding, an algorithmic approach*, Kluwer Academic Publishers, 433 pp., 1991.

[2] M. A. Arbib and A. R. Hanson, *Vision, brain, and cooperative computation*, MIT Press, 730 pp., 1987.

[3] R. E. Blahut, *Theory and practice of error control codes*, Addison-Wesley, 500 pp., 1983.

[4] R. J. Clarke, *Transform coding of images*, Academic Press, London, 432 pp., 1985.

[5] R. E. Crochiere and L. R. Rabiner, *Multirate digital signal processing*, Prentice-Hall, Signal Processing Series, 1983.

[6] A. Gersho and R. M. Gray, *Vector quantization and signal compression*, Kluwer Academic Publishers, 732 pp., 1992.

[7] R. M. Gray, *Entropy and information theory*, Springer-Verlag, 332 pp., 1990.

[8] E. C. Hildreth, *The measurement of visual motion*, MIT Press, Cambridge, 1983.

[9] B. Horn, *Robot vision*, MIT Press, 1986.

[10] A. K. Jain, "Image data compression: a review", *Proceedings of the IEEE*, Vol. 69, No. 3, pp. 349-389, March 1981.

[11] A. K. Jain, *Fundamentals of digital image processing*, Prentice-Hall information and system sciences series, 570 pp., 1989.

[12] N. S. Jayant and P. Noll, *Digital coding of waveforms: principles and applications to speech and video*, Prentice-Hall signal processing series, 688 pp., 1984.

[13] C. Labit, *Adaptativité et schéma de compression de données. Application au codage de séquences d'images télévisuelles*, Thèse de doctorat d'état, Université de Rennes-I, Feb. 1988.

[14] A. A. Liff, *Color and black and white: television theory and servicing*, 2nd Edition, Englewood Cliffs, 748pp., 1985.

[15] D. Marr, *Vision*, W. H. Freeman, New-York, 1982.

[16] A. N. Netravali and J. O. Limb, "Picture coding: a review", *Proceedings of the IEEE*, Vol. 68, No. 3, pp. 366-406, 1980.

[17] A. N. Netravali and B. G. Haskell, *Digital pictures: representation and compression*, Plenum Press, series: Applications of Communication Theory, 586 pp., 1988.

[18] W. W. Peterson, *Error correcting codes*, MIT Press, 1961.

[19] W. K. Pratt, *Digital image processing*, 2nd edition, John Wiley, 698 pp., 1991.

[20] M. Rabbani and P. W. Jones, *Digital image compression techniques*, SPIE Tutorial Texts in optical engineering, Vol. TT7, 220 pp., 1991.

[21] A. Singh, *Optical flow computation: a unified perspective*, IEEE Computer Society Press, 242 pp., 1991.

[22] S. Ullman, *The interpretation of visual motion*, MIT Press, Series in artificial intelligence, 229 pp., 1979.

[23] J. Woods, *Subband image coding*, Kluwer Academic Publishers, 355 pp., 1991.