PERCEPTION OF TIME-VARYING IMAGES

1 Introduction

1.1 Nature of perceived information

The previous chapter, by specifying the constraints linked to the definition of an image sequence digital signal, described the two essential procedures which make it possible to link information of a continuous nature both in time, space and spectral frequency, of perceived radiosities in a natural environment (3-D + t) and digital image sequence information $I(i, j; k)$; as a reminder, these two procedures are:

- a geometric projection operation on a spatiotemporal image plane (2-D+t)
- a sampling-quantization operation and spectral decomposition into color components, in order to characterize the perceived information as a finite set of discrete and quantized values (luminance, chrominance), localized in space and time, in discrete positions, sampled periodically (spatiotemporal pixels).

In this chapter, we will attempt to summarize the principal observations which the researchers were able to make, both in psychovision and in artificial vision theory, concerning the perception of moving scenes. It seems clear that these observations will depend on the nature of the observed signal (in terms of contrast, bandlimited spectrum, motion , etc...) but above all on the perceptual system itself: is it an image sensor (TV camera) in the sense of artificial vision, is it a complete visual system in the sense of human vision?

1.2 Nature of the observer

A simple analogy could be to consider that each eye of the human visual system "works" like a camera. This analogy can only be drawn when carrying out a very coarse structural and functional decomposition, in terms of:

- sensor input optical systems
- receiver unit surface: retinal surface for the visual system, cell matrix for a CCD sensor
- image information converter and transmitter to an interpretation system.

This simple analogy ceases when a more careful analysis of the biological or physio-chemical mechanisms is carried out. This chapter being devoted essentially to the perception of image information by the human visual system, below we emphasize more particularly the characteristics of perception within this framework. A summary of the main concepts for Human Visual System anatomical modelling is introduced in Appendix 2A. It is essential at this stage to make three comments:

1. Like all physical systems, the mechanisms of visual perception have their own physical limitations: *e.g.*, bounds of the visible spectrum, thresholds of sensitivity to contrasts, duration of temporal integration, effects of saturation (non-linearities) and of spatio-temporal filtering; this point will be discussed in more detail later on.

   As early as 1966, Robson [31] was able to demonstrate and measure the physical limits of the visual system by observation of the response curves of elementary receiver cells to variable temporal frequency stimuli and the resulting functions of sensitivity to contrast. We also note, from this first criterion, that the eye-sensor analogy is no longer possible. The orders of magnitude of the physical limits in both of these areas of perception are different and rarely comparable.

2. Contrary to a TV camera, in which the functional elements - information capture, digitalization (analog/digital converter for example), transport - are easy to isolate, this is not the case in a complete visual system. Certainly the eyeball plays the role of the optical system; the retinal cells are identifiable with the image acquisition cells; the photochemical mechanisms make it possible to transform the luminous radiosity information into electrical excitation; finally the optical nerve fibers make it possible to transfer the information to the interpretation area of the brain. However, all these biological elements are interdependent and operate continually in time (even if the temporal transmission band can be limited, as discussed in Sections 2.3 and 2.4) and almost continually in space since several million highly physio-chemically interactive nerve cells (spatial recovery) are involved in this case.

   The essential difference which we must note here therefore concerns the continuous perception of motion by our visual system. A number of authors [7] drew attention, supported by psycho-visual experiments, to the distinction which it is advisable to make between this continuous visual perception during time and a camera-sensor with a long exposure period. In the first case, a spatio-temporal filtering will generate an interpretation of motion. In the second, a blur caused by motion disturbs the acquired image information.

3. Finally, contrary to an image sensor providing an output raw image signal without interpretation (for example: output of a CCIR Rec.601 color signal), our visual system links together almost indissociably the information capture, transport and interpretation phases. Thus when studies are carried out into psychovision, the experiments and measures result from all of these basic information processing phases including, therefore the interpretation phase. Thus the elementary "visual" information conveyed by an optical nerve fiber will be processed, mixed and interpreted in the cerebral cortex (see perception models in Section 2.2) along with all the other temporally and spatially related information. This is also the case for motion perception.

1.3 Aims of studies into perception

Due to the enormous complexity of the structure of the brain and to the diverse nature of the mechanisms in question, studies concerning perception are essentially interdisciplinary, using concepts, models and experimental protocols concerning cellular biology, neurology, psycho-physiology and image signal processing (particularly concerning concepts of discrete linear systems and convolutional filters). There is, therefore, abundant but heterogeneous literature on the subject. This chapter can only hope to summarize some works carried out in the field. Essentially, we are looking to find out what models can be learnt from Perception Theory which may be useful in Image Processing. The image sequences coding cannot ignore these results since, generally speaking, the ultimate evaluation "system" of an encoding-decoding scheme will be a human visual system (for example that of the TV viewer) which will evaluate (perception and interpretation) the subjective quality of reconstituted image information. Studies into psycho-vision concerning Perception attempt to define the following concepts:

- descriptive models of perceived information: it could be frequency models, structural models or statistical models (fairly rare).
- geometrical models of acquisition optics: they are very similar to those handled in Artificial Vision; one of the most common approaches consists of adopting a perspective projection model.
- transfer function models (*e.g.*, convolutional filter type) for the elementary receptive cells of the retina and for the sensitive receptive fields of the cerebral cortex.
- spatial and temporal sensitivity models: the concept of such models is based on visibility threshold measurements [19], [35], [40]. In this way, assumptions are made concerning spatial separability and linearity, even if they are only approximations of the real operation of the visual system; this makes it possible to introduce simple experimental protocols; for example (Figure 1), a temporal sensitivity measurement on a 1-D sine wave pattern is carried out by modulating it by a sine wave attenuation function at a given temporal frequency,

  $$I(x,y;t) = I_0(1 + m\cos 2\pi ux \cos 2\pi wt) \tag{1}$$

  where I_0 measures the mean intensity, m the contrast of the sine wave pattern, and u and w the spatial and temporal frequencies tested respectively.
- models of masking effects: It is admitted [20], [21], [32], from both a spatial and temporal sensitivity point of view, that a given stimulus will generate different visual information depending on the context (masks) in which it occurs. This fundamental notion for image encoding is already used in the intra-image case to define adaptive methods for the choice of predictors or quantizers adapted to the context. Taking it into account as far as motion perception is concerned is less controlled.

The objectives described above, concerning research into perception, were mentioned explicitly by analogy with similar concepts defined in Artificial Vision. Other typologies are, of course, possible (see [24], [28]).

1.4 Perception and motion

These studies, though recent, began as early as 1960 (see Robson's work [31] on measuring the spatio-temporal sensitivity function) through, for example, the study of the temporal flicker of television tubes or stroboscopic effects of motion inversion (the link between temporal frequency of sampling and temporal frequency of motion) during a movie projection. They depended on models and theories already introduced at the level of static scenes such as:

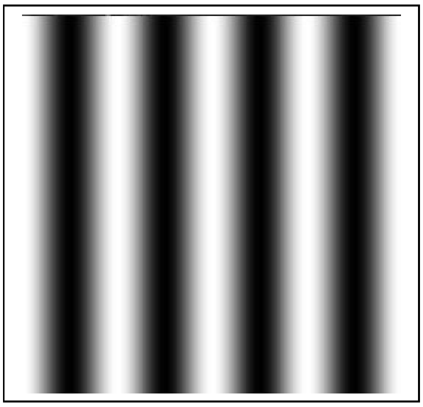


Figure 1: Spatiotemporal test gratings

- neuronal representation by receptive fields,
- low-level processings based to selective filters (see [24] and [42]),
- higher level processing by perceptual groupings towards interpretation tasks (see [44]).

Two scientific communities make complementary contributions in this area. The first, in the field of psychophysics, gives results in terms of sensitivity to various stimuli (contrast, motion) an excellent review is given by Burr ([6] and [7]); the second, based on the linear systems theory, models the receptive fields of the visual cortex in terms of adapted linear filters, selective from a point of view of spatio-temporal orientation. Watson [42] describes the general framework of such a model. Adelson [1] provides an explicit model of such filters adapted to motion based on the measurement of spatio-temporal energy in several spatio-temporal frequency bands. Heeger [14] demonstrates the possible link between these two approaches and that with algorithmic processes for apparent motion estimation (see Chapter 3).

1.5 Intermediate conclusion

In this introduction we wanted to evoke the possible differences and analogies between Human and Artificial Vision and between their respective sensors. As previously mentioned, all studies into perception by the human visual system are extremely important for the image sequences coding. In this respect we are interested, more especially, in Motion Perception (Sections 2.3 and 2.4) and wish beforehand to introduce the principal systematic tools required for such a study (Section 2.2). A few remarks concerning the implications for the analysis of motion and the image coding will be given in Section 2.5.

2 Modelling aspects

2.1 Notion of receptive fields

Much work, including that of D. Marr [24], has demonstrated that our visual system allows us to be sensitive to some features of objects viewed, such as: object boundaries or occlusion contours, corners (points of maximum curvature), regions (homogeneous textural areas).

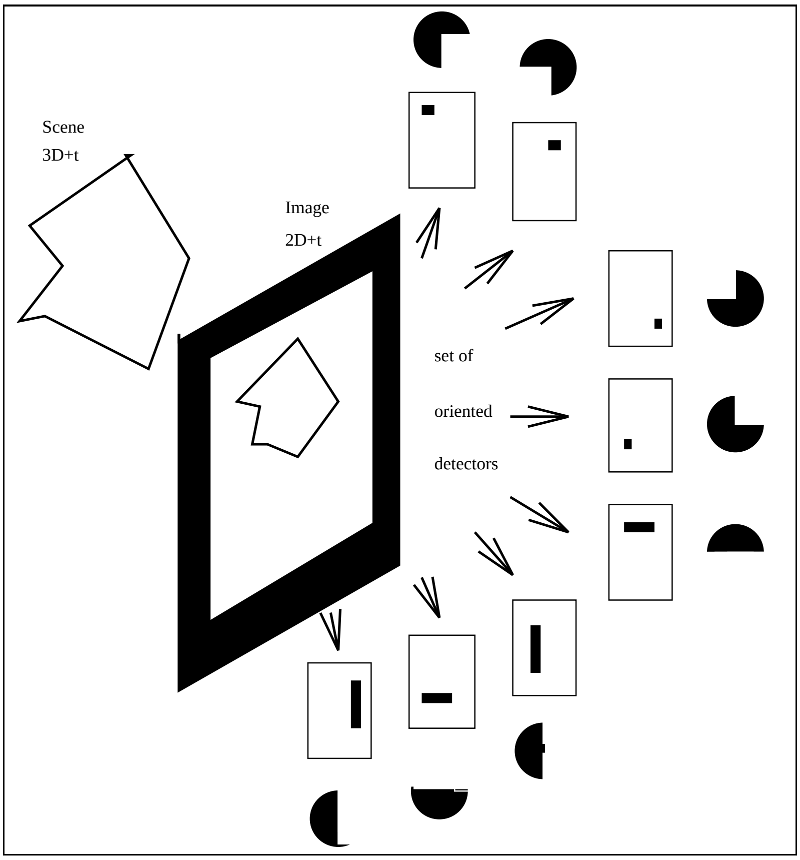


Figure 2: Multiple oriented feature detectors (From [16])

S. Ullman [38] expands this concept further to apparent motion. This observation is similar to that made by psychophysicians concerning the existence of receptive fields which, using excitation and inhibition of elementary detectors, makes it possible to diagnose the presence (or absence) of an oriented feature (see Figure 2).

For a visual scene containing "corner" and "edge" type characteristics with varying orientations, polarities and contrasts, rows of specific detectors make it possible to detect these characteristics.

2.2 The vision-interpretation link

There are three conventional modelling levels:

1. **the elementary cell**: this has a receptive field which is selective in character. If an oriented stationary stimulus is injected into it (for example a test pattern oriented to a given frequency), then areas excited and inhibited relative to a spontaneous activity of the unstimulated cell gives the detector of an oriented visual feature.

2. **the tuned column** composed of elementary cells all tuned to detect the same feature. The notion of tuning is the same as that introduced in a band-pass filter which models well the behaviour of an elementary receptive cell (see paragraph 2.2.3).

3. **the hypercolumn**, introduced in neuro-physiology, is composed of a tuned column set which makes it possible to apprehend all oriented stimuli selectively.

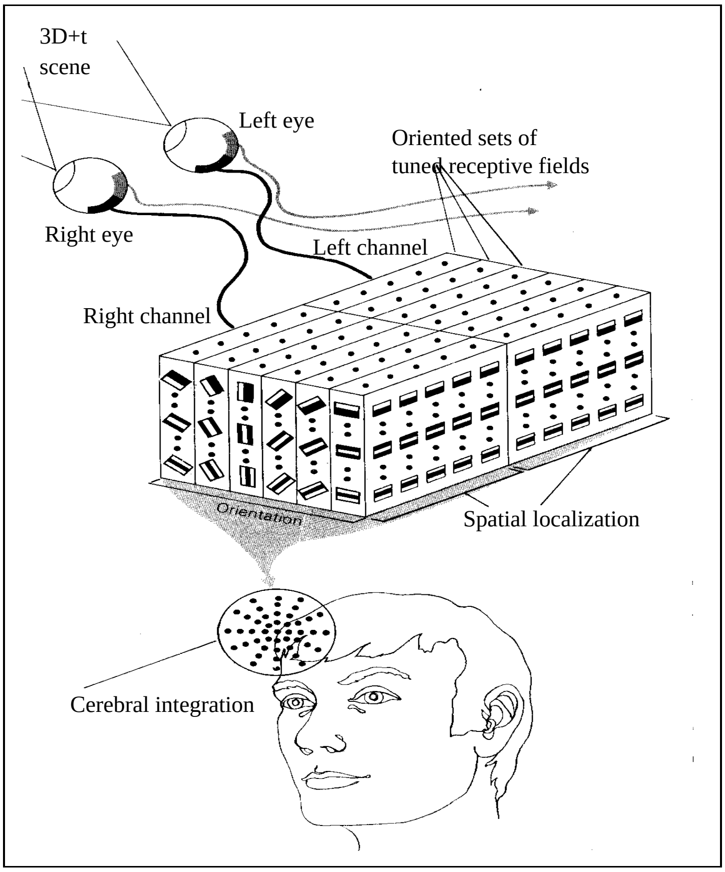


Figure 3: Example of modelling by tuned hypercolumns. Each tuned column is selectively oriented in spatial direction (orientation orthogonal to that of the receptive field) (From [16])

These three modelling levels make it possible to transfer visual informations from the elementary perception unit to the elementary processing unit for interpretation. Other more complex modelling levels (especially at the levels of elementary cells) can also be introducted.

2.3 Mathematical modelling tools

As previously mentioned, the theory of linear systems provides a good mathematical framework for modelling the "transfer function" of the human visual system and for expressing analytically the response to a given stimulus. Early several studies proposed the definition of the human visual system as a spatiotemporal linear filter. De Lange, Kelly, Roofs and Watson are amongst many authors who explored this type of modelling. Watson [42] thus proposes an analytical model for the filtering impulse response as depicted in Figure 4 and summarized as follows:

$$h(t) = \xi(h_1(t) - \zeta h_2(t)) \qquad (2)$$

with,

- $h_1(t) = u(t)(\tau(n_1-1)!)^{-1}(\frac{t}{\tau})^{n_1-1}\exp\left(-\frac{t}{\tau}\right)$ whose Fourier transform is $H_1(f) = (2i\pi f\tau + 1)^{-n_1}$. This is the response of a cascad of bandpass filters.
- $h_2(t)$ is identical filter with a different time constant $K\tau$ and length n_2.
- ξ is an amplitude gain factor.
- ζ is a weighting term called "transience factor" allowing the introduction of an attenuation at low frequencies.

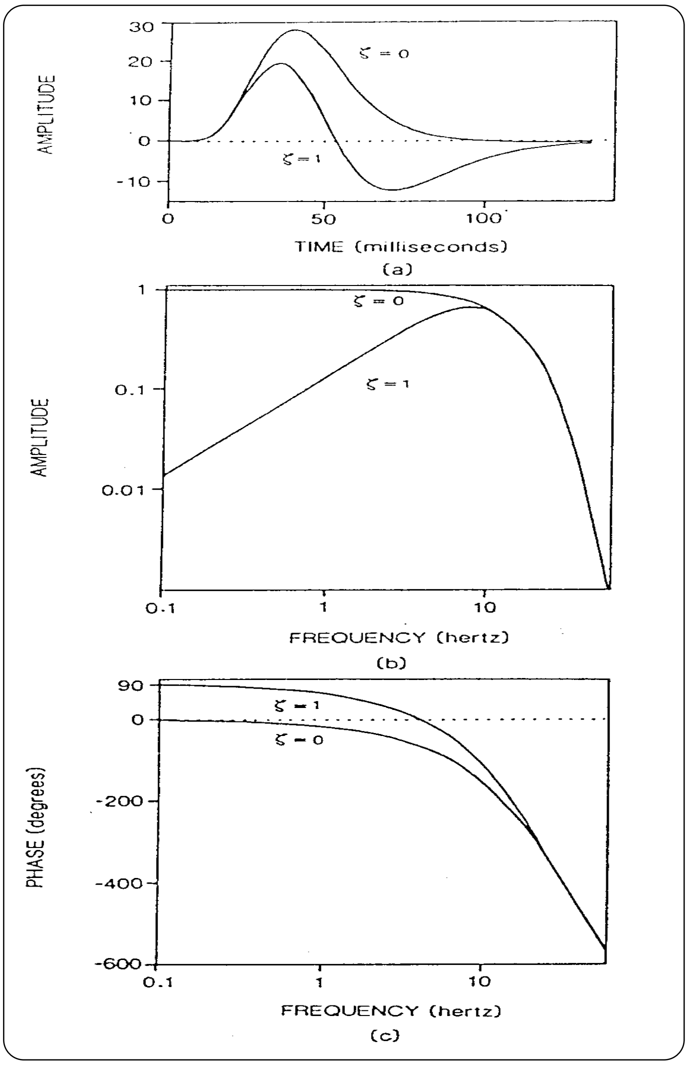


Figure 4: Human temporal sensitivity (From [42])

Other models exist, but all have the same aims:

- the definition of a visual system response curve is relative to a given stimulus
- the identification of **rapid** and **slow motions** often uses filters having impulse responses of longer or shorter duration

- the identification of **sustained** or **transitory** responses corresponds to detection filters. To a stationary stimulus, the first class of filters will provide a stationary response as a background task; the second class of filters will only provide a localized response at the moment the stimulus appears.

2.4 Experimental tools for tests of psychovision

Psychovisual tests are essentially based on the measurement of contrast visibility thresholds between

- a background region, which will consist of $I_B(x,y;t)$ information which is either constant (the case of stationary observation of a stimulus) or variable (study into the masking of a stimulus by a masking function to be identified).
- a "target" or "test" area into which the stimulus to be analyzed is injected ($I_T(x,y;t)$) provided with a rigid displacement $\vec{v}$ (see Figure 1).

Often an assumption of separability is introduced which makes it possible to define elementary stimuli. The notion of contrast and/or of visibility threshold when one of the parameters involved in the definition of $I_B(x,y;t)$, $I_T(x,y;t)$ or $\vec{v}(t)$ varies, is measured in accordance with the subjective test protocols defined, notably, by CCIR Rec.500 (see Appendix 2B). According to separability theory, spatio-temporal contrast can be expressed in the form

$$C(x,y;t) = \frac{I_T(x,y;t)}{I_B(x,y;t)} = C_{x,y}(x,y)C_t(t) \tag{3}$$

where,

$$I(x,y;t) = I_B(x,y;t) + I_T(x,y;t) = I_B(x,y;t)(1 + C(x,y;t)) \tag{4}$$

A simple test example is given to us by the following experiment [31]

$$I(x,y;t) = I_0(1 + m\cos 2\pi\omega_x x \cos 2\pi\omega_t t) \tag{5}$$

These measurements of contrast or visibility thresholds use very specific experimental protocols and particular units of measurement. We refer to certain elements of these in Appendices 2B and 2C of this chapter.

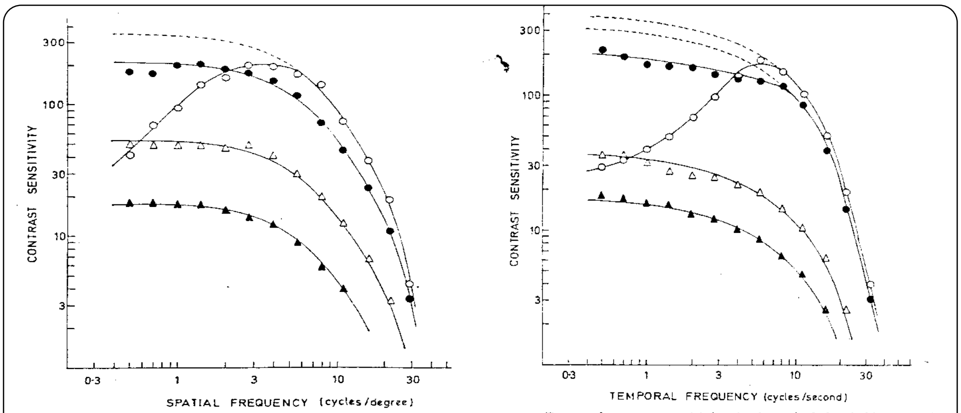
3 Principal results in the visual analysis of motion

The results mentioned here come from work carried out over a long period by psychophysicians concerning Perception and Motion [3], [5], [24], [26]. Many experimental results obtained from psychovisual tests on observers make it possible to formulate several assumptions concerning internal models of perception by the visual system in relation to motion and confirm the validity of mathematical models - in a restricted spatial and temporal frequency band - detailed in Section 2.4.

The starting point of all these experiments is certainly the observation of the interdependence of spatial and temporal sensitivity functions. Robson [31] provides the two principal conclusions of these measurements carried out on simple sine-wave patterns:

- at the high spatial (temporal resp.) frequency level, the cut-off in sensitivity is almost independent of the temporal (spatial resp.) frequency generated

- a lack of sensitivity at low spatial (temporal resp.) frequency levels is only observed in the case in which a low temporal (spatial resp.) frequency has also been generated. This distinct behaviour at low frequency, compared with the medium and high frequency bands, in part contradicts the theory of total separability of the mechanisms of perception of spatial and temporal stimuli. This theory is clearly inapplicable at low frequencies and remains slanted, by a gain in distinct sensitivity, to the other frequency bands.



Spatial contrast-sensitivity (reciprocal of threshold contrast) functions for different temporal frequencies. The points are the means of four measurements and the curves (one with a dashed low-frequency section) differ only in their positions along the contrast-sensitivity scale ○ 1 cycle per second, ● 6, △ 16, ▲ 22 cycles per second.

Temporal contrast-sensitivity (reciprocal of threshold contrast) functions for different spatial frequencies. The points are the means of four measurements and the curves (two with dashed low-frequency sections) differ only in their positions along the contrast-sensitivity scale, ○ 0.5 cycle per degree, ● 4, △ 16, ▲ 22 cycles per degree.

Figure 5: Spatial and temporal contrast sensitivity laws (From [31])

3.1 Sensitivity to motion

Robson's work in 1966 [31] was reformulated and complemented by Burr [6] making it possible to measure the sensitivity to motion of the perception system. In fact, on the basis of a simple sine-wave pattern of spatial and temporal frequencies ω_x and ω_t, its motion is then defined by (see also paragraph 2.4.3):

$$v = \frac{\omega_t}{\omega_x} \tag{6}$$

Figure 6, which is an extract of [6], shows sensitivity to motion; two essential remarks can be drawn from it:

1. Whatever the amplitude of the motion (here expressed in deg/sec), the bandwidth of the system remains roughly constant, as does the peak sensitivity.

2. The presence of large motions improves the visibility of the spatial low-frequency components.

3.2 Spatial and temporal tuning

In accordance with the analytical models of selective filtering for motion detection, psychovisual research shows that the human visual system has receptors sensitive to different spatio-temporal frequency bands:

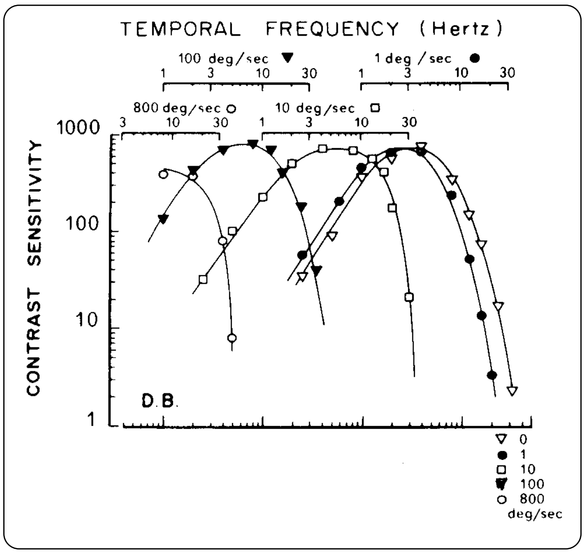


Figure 6: Sensitivity to motion (From [6])

Experiment 1: For a given contrast mask (experiments carried out show that the results in terms of relative sensitivity are linearly dependent on this factor, whatever the mask and pattern frequencies generated), the measurement of the increase of the visibility threshold of the pattern with or without mask is carried out:

- for various spatial and temporal frequencies
- for various pure translation motions (v)

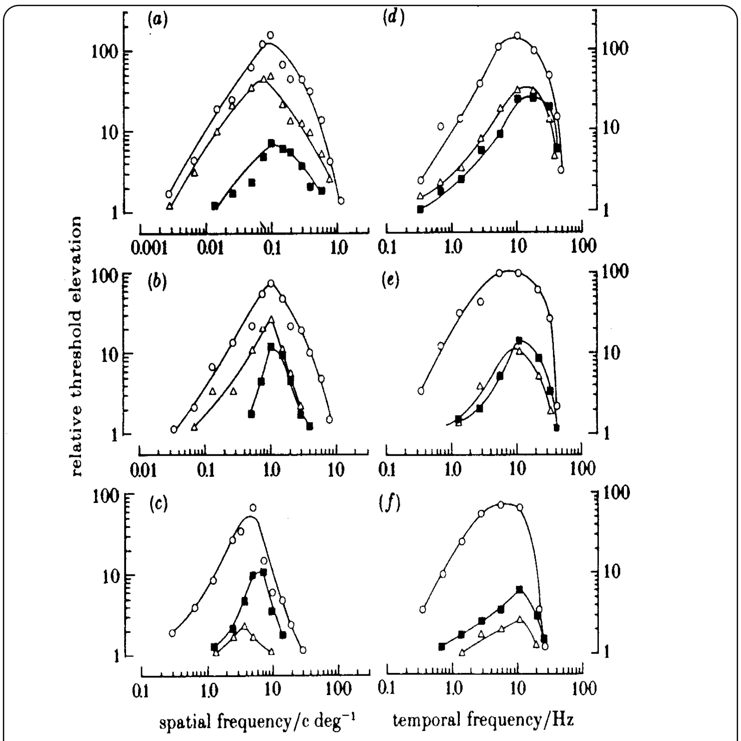
For each psychovisual test, we observe an extremum in the mask function when the mask spatial and temporal frequency coincides with those of the test stimulus. Furthermore, this mask function decreases strongly as soon as we move away from this "tuned" position. Each test corresponds to the excitation of a particular receptive field tuned on a particular frequency band (ω_x, ω_t).

Experiment 2: It is equally possible to observe these mask functions by varying the translation motion, v, of the test grating. One of the essential results visualized in the Figure 8 shows that the nature of the bandwiths observed depends on the motion parameter. Thus for an almost stationary grating, no attenuation of the low temporal frequencies is observed for the masking function. It is thus possible to conclude, as suggested by Tolhurst as early as 1973, that two types of detector exist:

- low-pass temporal frequency detectors which are not excited (or remain inhibited) in the presence of motion.
- detectors depending on motion which are both temporally and spatially band-pass.

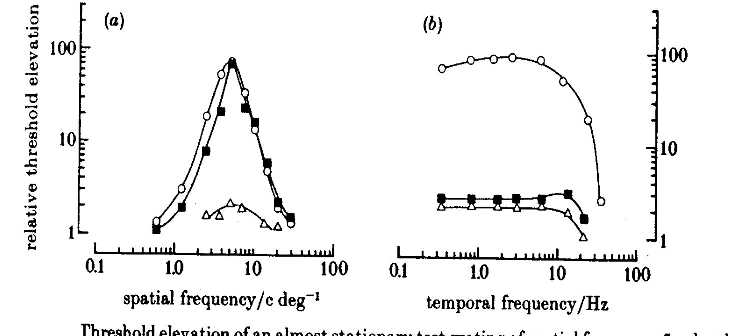
Experimental concept of adapted filters:

Burr [7] experimentally constructed the frequency responses of filters adapted to each motion amplitude of the test grating (remember that $v = \frac{\omega_t}{\omega_x}$). According to inverse Fourier transformation and to the theory concerning filter phases (linear phase or minimum phase filter) "receptive fields" or spatio-temporal impulse responses can be constructed (see Figure 9). Observation of these responses shows very clearly the adequation between psychovisual mechanisms and mathematical models by adapted filters.



Threshold expressed as the ratio of thresholds measured with the mask to that measured without it, for a few representative spatial and temporal mask frequencies (for S.J.A.). The test gratings had spatial frequencies of 0.1 (*a*, *d*), 1 (*b*, *e*) and 5 (*c*, *f*) c deg^{-1}, and were caused to drift at 8 Hz. For (*a*) the temporal frequencies were 2.8 (■), 11 (△) and 24 (○) Hz, and for (*b*) and (*c*) 2.8 (■), 11 (△) and 32 (○) Hz. For (*d*)–(*f*) the spatial frequencies were 0.25*f*, (■) *f* (△) and 4*f* (○), where *f* is the spatial frequency of the test.

Figure 7: Relative threshold elevation with and without spatiotemporal frequency masks (From [7])



Threshold elevation of an almost stationary test grating of spatial frequency 5 c deg^{-1} and drift frequency 0.3 Hz (velocity 0.06° s^{-1}), for a few temporal and spatial frequencies of the mask. The temporal frequencies reported in (*a*) are 0.7 (○), 11 (■) and 32 (△) Hz. The spatial frequencies in (*b*) are 1.25 (■), 5 (○) and 20 (△) c deg^{-1}.

Figure 8: Relative threshold elevation for different spatiotemporal frequencies of the mask (From [7])

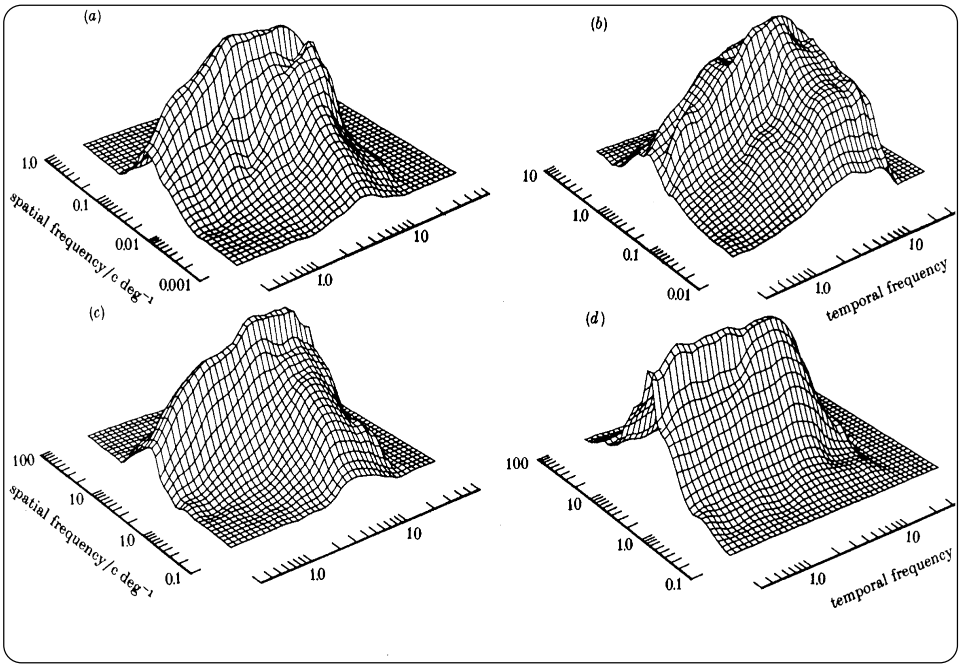


Figure 9: Spatiotemporal filters for 4 masking sets of test spatial and temporal frequencies (From [7])

3.3 Saccadic vision

All the previous experiments assumed the choice of a continuous perception modelling. What happens to sensitivity thresholds during saccadic motions of our oculomotor system? This problem also arises in the presence of motions with very large amplitudes.

Burr [7] described experiments into the behaviour of the human perception system in the presence of saccadic motions. The results of subjective tests clearly show the loss of sensitivity in the presence of saccadic motion, a loss which reduces as the tested spatial frequency increases, and which disappears once a certain frequency is reached.

The explanation of the underlying psychovisual mechanism shows that, in the case of saccadic perceptions (or, in duality, objects visualized by brief flashes (around 20 msec)) the receptive fields tuned to motion in this case become inactive. At a higher order, the same reasoning can also be applied when abrupt variations occur on a trajectory. An abrupt increment (or decrement) of the motion on a constant motion grating, can only just be seen in the presence of saccadic vision.

4 Motion perception by selective filters

4.1 Continuous/discrete approaches

Adelson and Bergen [1] proposed comparing continuous approaches with discrete ones in order to design motion perception; this modelling taxinomy is usual in the algorithmic domain (See Chapter 3: comparison between correspondence methods and differential methods).

The motion perception discrete approach tries to match isolated primitives (in the visual characteristics sense as introduced in paragraph 2.2.1). This is only realistic for single and overall motions, and probably unefficient for the processing (here in the sense of perception) of high spatio-temporal frequency motions.

The continuous approach, favoured by a number of authors, including Adelson [1], Watson [42] and Heeger [14], uses the concept of spatio-temporal filtering making it possible:

- to be compatible with the spatial perception models;
- to include, in the same modelling framework, the preprocessing of the acquisition of visual information (transfer function of the optical and nervous system) and the processing of perception/interpretation (filters adapted to spatio-temporal characteristics);
- to use known modelling tools of linear systems.

We also favour this filtering approach below, the references in the first case being able to be found in overall correlation and multi-image vision based models (stereo vision for example).

4.2 Short and long-term visual mechanisms

Several studies demonstrated the existence of two mechanisms with different characteristics concerning the perception of motions which are very localized in time ("short-range motions") and more long-term motions ("long-range motions") which can have something in common with notions of trajectories.

The perception of short-term motions concerns low-level visual information, only perceptible during the rapid presentation of moving stimuli. On the other hand, the perception of long-term motions, like the perception of spatial stimuli, brings into play a notion of spatio-temporal perceptual grouping.

This type of separation can occur quite naturally in a system of modelling by filter banks and decomposition into frequency subbands which, by nature, would be asymmetric along the temporal frequency axis. A higher level modelling stage nevertheless remains necessary in order to deal with the case of complex trajectories.

4.3 Motion and frequency decomposition

It is important to stress the intrinsic relationship which links motion (apparent projected in the image plane) and frequencies.

In the case of a 1-D signal (an image line for example) decomposed into spatial frequencies indexed by ω_x and motion v, we get the following relationship

$$v = \frac{\omega_t}{\omega_x} \tag{7}$$

where ω_x is an elementary frequency of the 1-D signal expressed in cycles/pixel, ω_t the associated temporal frequency expressed in cycles/s and v expressed in pixels/s.

The image signal is assumed to be sampled in time; all signals composed of temporal frequencies greater than the Nyquist frequency (1/2 cycle/frame) will be degraded by spectral overlap. This gives a maximum limit for motion v of all elementary frequencies

ω_x defined by the Equation (7). Generally speaking, the temporal spectrum of a 1-D signal in constant motion v is a linear function (slope $= v$) of the spatial spectrum.

Extrapolation in the case of a 2-D image signal of spectral components (ω_x, ω_y) is evident. In this case:

$$\omega_t = u\omega_x + v\omega_y \tag{8}$$

if $(u, v)^t$ indicates the apparent 2-D motion of the signal.

The part of the spatio-temporal spectral space $(\omega_x, \omega_y, \omega_t)$ occupied by this object in motion is therefore strictly limited to one plane (if the object is in constant motion $(u, v)^t$).

At this level we see appearing the possibility of extracting motion parameters, as much from a visual as an algorithmic point of view, by frequency filtering techniques.

4.4 Motion and orientation

Observation of a single object (here a vertical bar) with constant motion is expressed, as much in the frequency domain (see previous paragraph) as in the spatio-temporal domain itself (x, y, t) by search of oriented-planes. Motion perception can therefore only be achieved through perceptive elements which are selectively oriented.

By analogy with the spatial case (see paragraph 2.2.2) where we have mentioned the possibility of modelling by receptive fields which are selectively oriented, here it is a question of extending this notion to the modelling of receptive fields oriented in the spatio-temporal domain (x, y, t) or in its frequency dual $(\omega_x, \omega_y, \omega_t)$. The motion perceived for an oriented receptive field will be in an orthogonal direction along the selective orientation of the field considered (see Figure 10).

The problem of modelling for motion perception leads then to the concept of oriented-filter banks. Often the theory of separability in space and time is used, which makes it possible to take up again the models defined for spatial perception and to add to them selective temporal filters for motion perception. Two experimental examples illustrate this point. As selective receptive field modelling filters, Heeger [14] uses a bank of 3-D Gabor filters whose impulse response is a sine-wave function modulated by a Gaussian

$$h(x, y, t) = \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x\sigma_y\sigma_t} e^{-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}+\frac{t^2}{2\sigma_t^2}\right)} \sin 2\pi(\omega_{x_0}x + \omega_{y_0}y + \omega_{t_0}t) \tag{9}$$

with which we can associate the bank of symmetrical cosine-curve filters and can define a Gabor 3-D energy filter, by a summation of energy responses from each filter. The resulting filter bank is of the type:

$$\begin{aligned} H(\omega_x, \omega_y, \omega_t) = {} & \frac{1}{4}e^{-4\pi^2\left(\sigma_x^2(\omega_x-\omega_{x_0})^2+\sigma_y^2(\omega_y-\omega_{y_0})^2+\sigma_t^2(\omega_t-\omega_{t_0})^2\right)} \\ & +\frac{1}{4}e^{-4\pi^2\left(\sigma_x^2(\omega_x+\omega_{x_0})^2+\sigma_y^2(\omega_y+\omega_{y_0})^2+\sigma_t^2(\omega_t+\omega_{t_0})^2\right)} \end{aligned} \tag{10}$$

that is to say, the sum of two 3-D Gaussian filters.

Thus it is possible to specify a parametric family of H filters, all tuned on the same spatial frequency band (*i.e.*, $\omega_{x_0}^2 + \omega_{y_0}^2 = \text{cte}$), but with selective spatial and temporal orientation and, therefore, with selective spatial and motion characteristics. The approach described in [14] illustrates a cylindrical segmentation of the spectral space $(\omega_x, \omega_y, \omega_t)$

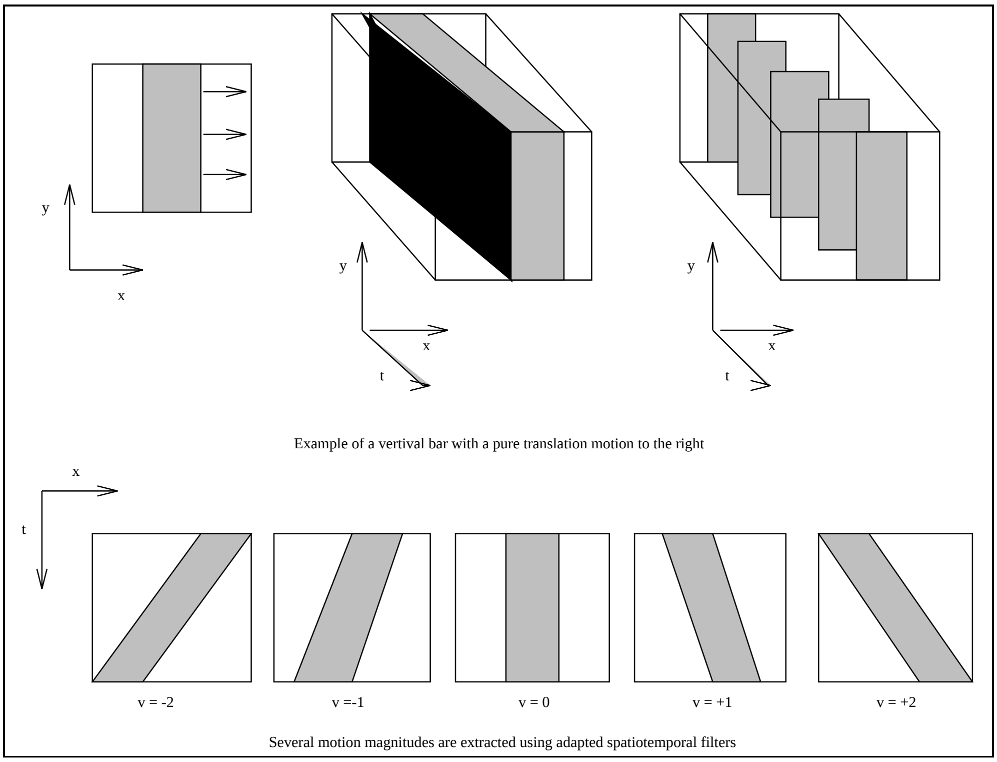

y
x
y
x
t
y
x
t
Example of a vertival bar with a pure translation motion to the right
x
t
v = -2
v =-1
v = 0
v = +1
v = +2
Several motion magnitudes are extracted using adapted spatiotemporal filters

Figure 10: Motion filtering using spatiotemporal oriented receptive fields (From [1])

making it possible to define oriented motion detectors tuned to a temporal frequency of ω_{t_0} (in 3 layers: "leftward" motion , stationary and "rightward" motion regions) and according to several spatial orientations.

As for Watson [41], he defined an elementary motion sensor using summation of 3-D separable filters into x, y and t by introduction of Gabor 2-D type spatial filters and a temporal filter such as that described in paragraph 2.2.3. This type of modelling leads to a usual shape (for the x direction) such that:

$$a(x)b(y)c(t)(\delta(x,y,t) + \delta(y)h(x)h(t)) \tag{11}$$

where $a(x), b(y)$ and $c(t)$ are separable filters selected for their close properties to human perception (*e.g.*, Gabor filters) and where the second term indicates the quadrature filter (by application of the Hilbert transform $h(x)$ and $h(t)$ [1])

4.5 Some examples of properties

Based on families of selective filters previously described, Adelson [1] proposes a form of complete modelling for a motion detector having the following properties:

- separability of spatial and temporal filters
- choice of two spatial filters $F1$ and $F2$, shifted in phase, and of two temporal filters with different frequency bands
- construction of oriented linear responses dependent on contrast

[1] As a reminder, $h(x) = -\frac{1}{\pi x} \Longleftrightarrow H(\omega_x) = i.\text{sign}(\omega_x)$

- measurement of the output spectral energy of selective bands with identical motion orientation

- relative energy measurement (zero value for stationary areas) for motions in opposite directions. The latter is due to the fact that within a single image zone, two motions in opposite directions cannot be perceived.

- apparent displacement measurement, from the detection of several elementary motion sensors.

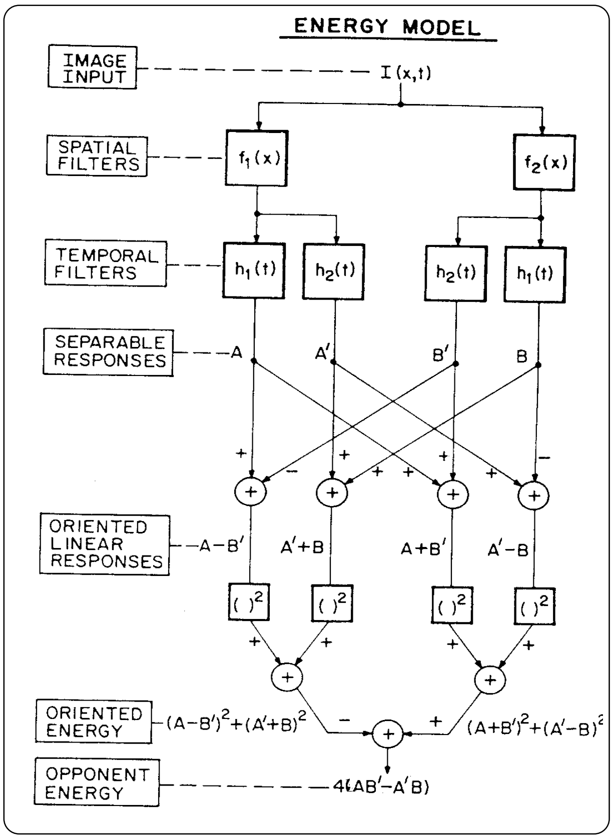


Figure 11: Selective mirror filter bank as motion detector (From [1])

This complete study of the modelling of motion perception tends to confirm the hypotheses in the field of psychovision, concerning the real mechanisms brought into operation within the visual system. The filtering procedure proposed in Figure 11, taken from [1] carries out the desired analysis of motion in different spatio-temporal frequency bands.

5 Motion perception and application to coding schemes

5.1 Subband coding

In Chapter 6 we will describe the methods of coding digital image sequences for the decomposition into frequency subbands [43], and the analysis of motion on the latter.

It seems to be attractive to make the connection between the effective methods from an algorithmic point of view of subband decomposition and models of adapted selective filters, such as defined for psychovisual criteria. It should be noted that this comparison is justified concerning the following points:

- separability of filters, notably of spatial filters in respect to temporal filters
- dissymmetric spectral partition in time and space. Thompson [36] (see Figure 12) or Burr [6] proposed "ideal" partitions of the spatiotemporal frequency space.
- design of energy-model filters by summation of outputs of quadrature mirror filters
- several families of filters which are close to psychovisual modelling are those which are also currently being tested in approaches to subband encoding. Let us quote, as a reminder: Gabor 3-D filters [41], [14], high-order derivative filters of Gaussian filters [24], [1], filters based on the Hermite transform [23].

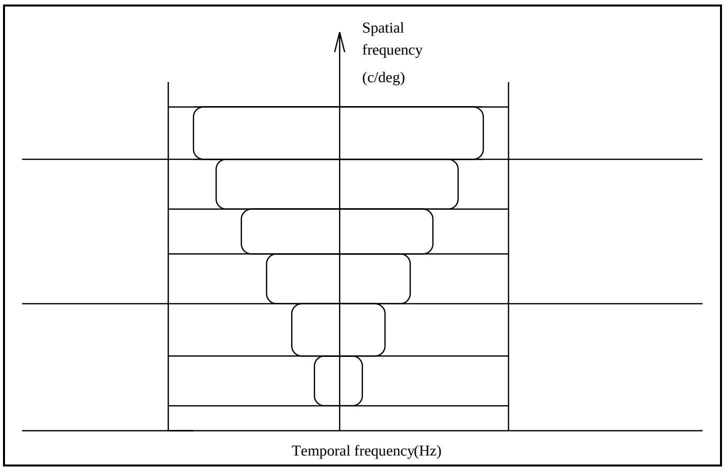


Figure 12: An ideal partition of the spatial/temporal frequency space (From [36])

The motion perception elements modelled here should make it possible, in real coding techniques, to choose the motion estimators (from the point of view of quantization and dynamic of estimated motions) and quantizers adapted to each frequency band for the encoding of prediction errors with motion compensation.

5.2 Motion perception and analysis

Given that the evaluation of a Motion Analysis System is very often carried out using the perception characteristics of a human visual system, observer of the sequence being analyzed, it appears necessary in the motion analysis algorithm to take into account several elements of psychovisual perception. In reading Chapter 3, it can be seen that this is not yet the case, at least in conventional motion estimation algorithms designed before the present results were achieved in dynamic psychovision. We can, however, note the typology similarity between motion analysis methods, in the areas of both algorithms and visual analysis:

1. low-level or local vision mechanisms adopt a continuous approach to the analysis of motion. In this context we again find filtering techniques and differential methods.
2. on the other hand, and complementary to this, the overall processing is carried out by perceptual grouping [44] and by matching discrete structures.

This methodological classification is also usual for motion analysis algorithms.

5.3 Spatio-temporal interpolation

This topic will be dealt with, from an algorithmic point of view, in Chapter 7. It is natural to think that the choice of temporal frequencies for sampling by an image sequence signal, is based on the spatio-temporal interpolation property of the visual system. This property has been researched more particularly in psychovision by Burr *et al.* [7], Fahle and Poggio [15], and Morgan and Watt [27]. In the case of experiments carried out by [15] in order to determine the spatial and temporal visual acuity thresholds on vernier targets, characterized by the following parameters:

$$\begin{aligned} \Delta x &= \text{spatial separation (offset)} \\ \Delta t &= \text{temporal separation} \\ v = \tfrac{\Delta x}{\Delta t} &\text{ constant motion of gratings} \end{aligned}$$

the following conclusions are presented: spatial separation acuity is constant across a wide range of velocities (from 0 deg/s to 4 deg/s); the temporal acuity, on the other hand, depends on the spatial separation (Δx) and speed generated. An optimal acuity is observed for each separation Δx : this decreases when Δx increases. Obviously, if a spatial graduation (by high frequency attenuation) is applied to the target, the spatial acuity increases. However, for large motions and wide spatial separations, the temporal acuity is improved. This result obtained in psychovision can be compared with the different observation made in algorithms for motion analysis where only the addition of a high-frequency noise makes it possible to improve the estimation results (for example by increasing the gradient estimates in differential methods).

Finally, a hybrid experiment is carried out concerning the mixing of spatial Δx and temporal offsets which do not necessarily obey the relationship of rigid motion ($v = \frac{\Delta x}{\Delta t}$). The essential conclusion drawn is that motion compensation (in the sense of spatial offset compensation by equivalent temporal offset around the precise velocity v) only occurs for a small spatial offset; in the other cases, the spatial separation mechanism overrides (see [15]).

In conclusion, these authors evoke physiological and functional reasoning of spatiotemporal interpolation which, in addition to increasing visual acuity and the accuracy of temporal and spatial localization, also makes it possible to introduce a notion of spatial and temporal continuity of the visual field, a theory often held in digital algorithmics.

6 Conclusion

In this chapter and in summarizing experiments that various researchers, often separately in the fields of Neurophysiology and Image Processing, have carried out over several decades, our aim was to recall that the motion perception models introduce mechanisms which are at the same time biological, neurological, psychological and algorithmic. An elegant modelling context which makes it possible to perceive these mechanisms, in a unified fashion, as well as that of spatio-temporal filters selective according to orientation and motion. We evoke deliberately and essentially this point of view and possible methodological comparisons.

Without doubt these models remain somewhat unprecise, too qualitative, and dependent on the observation conditions and on the observers (even if the nature of the sensitivity curves remains approximatively stable between several observers). Furthermore, the tests carried out into motion perception involve simple objects and scenes: sine-wave patterns tested against constant or mono-frequency backgrounds, constant pure translation motion for the moving object analyzed. These tests remain essential to the design of local perception models. The design of overall perception models of natural complex scenes, composed of several objects and various given motions, still remains at the research stage. Nevertheless, it is certain that in order to design effective methods of image sequence coding - the term effective should be taken in the sense of a close approximation of the perceptive system - an objective quality model on dynamic sequences will have to be designed. Such research into psychophysics [21], [32] resulted in the design of masking functions, introduced effectively into spatial prediction functions for example. The next stage is, of necessity, the introduction of such spatiotemporal masking functions into the processing modules of an encoding system dependent on motion, whether at the level of a predictor, quantizer or interpolator. Similarly, as far as frequency decomposition techniques are concerned, certain studies take account of some elements of perception in the selection of transform coefficient quantizer characteristics [30] (for example: quantization visually adapted for each DCT coefficient). These should be extended to the context of decomposition into spatiotemporal frequency subbands.

The assumption of constant velocity, systematically adopted in the context of this chapter, is obviously very restrictive: it would be advisable to overcome it by two methodological extensions:

- study of the visual response to a velocity step (*i.e.*, in the presence of a discontinuity in the velocity vector field, essentially localized around motion occlusion areas); this extension is the logical continuation of researches into perception in the presence of purely spatial discontinuities (contour masking);

- research into the visual response to a “natural” dynamic scene (*i.e.*, having several objects with velocities distinct by both amplitude and direction)

In conclusion, the studies into motion perception constitute an essential research challenge in better understanding the complex mechanisms, both biological and functional, of the human visual system. Despite a considerable number of tests and experimental measurements, these studies presently only give essentially qualitative results and a methodological guide for modelling visual receptive fields tuned to a particular motion characteristic. This research will certainly progress if we face up to and comfort ourselves with the results obtained in motion analysis from an algorithmic point of view, and in artificial vision.

7 Appendices

7.1 Appendix 2A: Anatomical model of the Human Visual System

The principal components of the Human Visual System can be summarized as follows (see [16], [24], [29] for further details):

- **Human eye**: This sensitive receptor is easily characterized by a perspective projection system where a 3-D object in the viewed scene is focussed by the **cornea** and **lens** to form a 2-D projected object on the **retina** which is located at the back of the eye ball and connected to the **optic nerve**.

- **Eye movements**: Eye movements can be voluntary or instinctive. These movements when they are voluntary and so controlled enable to track moving objects. However unvoluntary movements like slow drifts of the point of fixation, saccades, occur and degrade the analyzed image even if they ensure continuous visual activities over the retina.

- **Photoreceptive cells**: The retina is usually described as a set of photoreceptive cells generating the energy responses to the neural layers. These receptors are of two kinds: **rods** for low light vision, **cones** for color vision at normal daylight levels.

- **Neural synapses**: Several layers of neural synapses transform the photonic signal into electric potentials by chemical reactions. These potential are finally coded and transmitted on the nerve fibers towards the **visual cortex** for cerebral interpretation.

7.2 Appendix 2B: Subjective psychovisual tests: CCIR Recommendation 500

This CCIR recommendation proposes a methodological standard to perform subjective tests for evaluating quality measures between original and processed pictures. This evaluation method suggests different proposals to guide, the most objectively as possible, any test procedure.

- **choice of observers**: a minimum of ten different observers is recommended; they could be experts or not of the processing to be tested, even if not specialists would be better

- **scale for subjective evaluation**: a scale of 5 levels is recommended: from excellent quality with unperceptible degradation (level 5) to inappropriate quality and very annoying degradation (level 1)

- **scale for comparative tests**: a 7-level scale is proposed and enables to quantize highly positive improvement (level +3) to highly negative decrease (level -3) by the way of identical results (level 0) when comparative tests are performed; this relative scale is used, for example, when sequential tests try to evaluate the subjective perception of different sets of parameter for a given stimulus.

- **observation conditions**: a standardized set of parameters is given for the experimental conditions:

- observation distance ($4H$ or $6H$) relative to the screen height (H)
- peak-luminance of the display ($70 \pm 10cd/m^2$)
- ratio between peak-luminance and minimal luminance for an inactive display (< 0.02)
- ratio between background and maximal image luminances ($\simeq 0.01$)
- room illumination
- chromaticity of the background
- ratio between solid angles from the observer viewpoint and relative to the screen and to the viewed background (≥ 9)

Some other international reports (R313-4 and R405-3) give complementary details for these conditions.

7.3 Appendix 2C: Units of measurement used in psychophysics

- Luminous Energy Q: in *lumen-second* or *talbot*
- Radiant Flux Φ (Q/time): in *lumen*
- Intensity:
  - $\frac{\Phi}{\text{solid angle } \Delta\Omega}$: in *lumen/steradian* or *candela*
  - Illumination $\frac{\Phi}{\text{area } A}$: in *lumen/meter*^{2} or *lux*
  - Luminance $\frac{\Phi}{\text{area } A. \text{ solid angle} \Delta\Omega}$: in *candela/meter*^{2} or *lambert*
- Spatial grating frequency: in *cycles/deg*
- Temporal grating frequency: in *Hz* or *cycles/sec*
- Grating velocity: in *deg/sec*

References

[1] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion", *Journ. Optical Soc. America*, Vol. 2, No. 2, pp. 284-299, Feb. 1985.

[2] M. Arbib and A. R. Hanson, eds, *Vision, brain and cooperation computation*, MIT Press, Cambridge, 1987.

[3] S. M. Anstis, "Apparent Movement", in: *Handbook of sensory, Vol. VIII, Perception*, Springer-Verlag, 1977.

[4] S. M. Anstis, "Visual coding of position and motion", in: M. Arbib and A. R. Hanson, eds, *Vision, brain and cooperation computation*, MIT Press, Cambridge, 1987.

[5] N. I. Badler and J. K. Tsotsos, eds, *Motion representation and perception*, North-Holland, 1983.

[6] D. C. Burr and J. Ross, " Visual analysis during motion", in: M. Arbib and A. R. Hanson, eds, *Vision, brain and cooperation computation*, MIT Press, Cambridge, 1987.

[7] D. C. Burr, J. Ross and M. C. Morrone, "Seeing objects in motion", *Proc. of the Royal Society of London*, Series B 227, pp.249-265, 1986.

[8] P. J. Burt, "The interdependance of temporal and spatial information in early vision", in: M. Arbib and A. R. Hanson, eds, *Vision, brain and cooperation computation*, MIT Press, Cambridge, 1987.

[9] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of grating", *Journ. of Physiology*, Vol. 197, pp. 551-566, 1968.

[10] CCIR Recommendation 500, "Method for the subjective assessment of the quality of T.V. pictures", *CCIR 13th Plenary Session*, Geneva, 1974.

[11] B. Cohen and J. Bodis-Wollner, eds, *Vision and the brain: the organization of the central visual system*, Raven-Press, 1990.

[12] J. G. Daugman, "2-D spectral analysis of cortical receptive field profiles", *Vision Research*, Vol. 20, pp. 847-856, 1980.

[13] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2-D visual cortical filters", *Journ. Optical Soc. America*, Vol. 2, No. 7, pp. 1160-1169, July 1985.

[14] D. J. Heeger, "Model for the extraction of image flow", *Journ. Optical Soc. America*, Vol. 4, No. 8, pp. 1455-1470, Aug. 1987.

[15] M. Fahle and T. Poggio, "Visual hyperacuity spatiotemporal interpolation", *Proc. of the Royal Society of London*, Series B 213, pp. 451-477, 1981.

[16] J. Frisby, *Seeing*, Oxford University Press, 1979.

[17] B. Girod, "Psychovisal aspects of image communication", *Signal Processing*, Vol. 28, pp. 239-251, 1992.

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

[19] D. H. Kelly, "Motion and Vision: stabilized spatiotemporal threshold surface", *Journ. Optical Soc. America*, Vol. 69, pp. 1340-1349, 1979.

[20] G. E. Legge and J. M. Foley, "Contrast masking in human vision", *Journ. Optical Soc. America*, Vol. 70, pp. 1458-1471, 1980.

[21] S. R. Lehky, "Temporal properties of visual channels measured by masking", *Journ. Optical Soc. America*, Vol. 2, No. 8, pp. 1260-1272, Aug. 1985.

[22] F. X. Lukas and Z. L. Budrikis, "Picture quality prediction based on a visual model", *IEEE Transactions on Communications*, Vol. COM-30, No. 7, pp. 1679-1692, July 1982.

[23] J. B. Martens, “The Hermite transform: theory and applications”, *IEEE Transactions on Acooustics, Speech and Signal Processing*, Vol. ASSP-38, No. 9, pp. 1595-1618, Sept. 1990.

[24] D. Marr, *Vision*, Freeman, 1982.

[25] M. S. Marx and J. G. Mat, “The relationship between temporal integration and persistence”, *Vision Research*, Vol. 23, No. 10, pp. 1101-1106, 1983.

[26] M. Miyahara, “Analysis of perception of motion in Television signals and its application to bandwith compression”, *IEEE Transactions on Communications*, Vol. COM-23, pp. 761-768, July 1975.

[27] M. J. Morgan and R. J. Watt, “On the failure of spatiotemporal interpolation: a filtering model”, *Vision Research*, Vol. 23, No. 10, pp.997-1004, 1983.

[28] K. Nakyama, “Biological image motion processing: a review”, *Vision Research*, Vol. 75, pp. 626-660, 1985.

[29] A. N. Netravali and B. G. Haskell, *Digital pictures: representation and compression*, Plenum Press, 586 pp., 1988.

[30] P. Pirsch, “Design of DPCM quantizers for video signals using subjective tests”, *IEEE Transactions on Communications*, Vol. COM-29, No. 7, pp. 990-1000, July 1981.

[31] J. G. Robson, “Spatial and Temporal Contrast-Sensitivity Functions of the Visual System”, Vol. 56, pp. 1141-1142, Aug. 1966.

[32] B. E. Rogowitz, “Spatial/Temporal interactions: backward and forward metacontrast masking with sine-wave gratings”, *Vision Research*, Vol. 23, No. 10, pp. 1057-1073, 1983.

[33] B. Sakitt and H. B. Barlow, “A model for the economical encoding of the visual image in cerebral cortex”, *Biol. Cybernetics*, Vol. 43, pp. 87-108, 1982.

[34] D. J. Sakrison, “On the role of the observer and a distorsion measure in image transmission”, *IEEE Transactions on Communications*, Vol. COM-25, No. 11, pp. 1251-1267, Nov. 1977.

[35] P. Thompson, “Discrimination of moving gratings at and above detection threshold”, *Vision Research*, Vol. 23, pp. 1533-1538, 1983.

[36] P. Thompson, “The coding of velocity of movement in the human visual system”, *Vision Research*, Vol. 24, No. 1, pp. 41-45, 1984.

[37] D. J. Tolhurst, “Separate channels for the analysis of the shape and the movement of a moving visual stimulua”, *Journ. of Physiology*, Vol. 231, pp. 385-402, 1973.

[38] S. Ullman, *The interpretation of visual motion*, MIT Press, Cambridge, 1979.

[39] S. Ullman and E. C. Hildreth, “The measurement of visual motion”, in: O. J. Braddick and A. C. Sleigh, eds, *Physical and Biological Processing of Images*, Springer-Verlag, 1982.

[40] A. B. Watson and J. C. Robson, "Discrimination at threshold: labelled detectors in human vision", *Vision Research*, Vol. 21, pp. 1115-1172, 1981.

[41] A. B. Watson and A. J. Ahumada, "Model of human visual-motion sensing", *Journ. Optical Soc. America*, Vol. 2, No. 2, pp. 322-340, Feb. 1985.

[42] A. B. Watson, "Temporal Sensitivity", in: *Handbook of perception and human performance*, 1986.

[43] J. Woods, ed., *Subband image coding*, Kluwer Academic Publishers, 1991.

[44] S. W. Zucker, "The diversity of perceptual grouping", in: M. Arbib and A. R. Hanson, eds, *Vision, brain and cooperation computation*, MIT Press, Cambridge, 1987.