Motion and Structure. Application to feature-oriented coding

As mentioned in Chapters 1 and 3, observation and associated estimation methods of an apparent motion vector field within an image sequence results from the projection of 3-D objects and from their 3-D motion on the 2-D image plane. This projection operation, which is perspective or orthogonal in nature, depending on the projection system selected, creates ambiguities concerning apparent 2-D motions perceived and, in addition, does not generate a compact representation of the motion information itself. In fact, if we take the example of a rigid 3-D body undergoing 3-D motion, this motion of the object is wholly specified by a small number of parameters (generally six degrees of freedom) through the kinematic screw (translation + rotation) associated with the object and referenced in relation to an absolute fixed reference. This same 3-D motion observed through the 2-D apparent motion-vector field is, on the other hand, much more complex to analyze and to represent. A more compact representation and more effective estimation of complex motions which are not purely translational parallel to the image plane, constitute the two essential arguments in favour of a higher level modelling for the motions and structures of objects manipulated. All of the motion estimation techniques detailed in the preceding chapters limited themselves:

- to a local estimation by pixel for which the representation of motion by its apparent motion vector $(\dot{x}, \dot{y})^t = (\frac{dx}{dt}, \frac{dy}{dt})^t = (u, v)^t$ - two translational components - is adequate. Clearly, it is impossible to talk about rotational motion of an object restricted to one pixel.

- to a global estimation of a translation vector $(u, v)^t$ by block (block matching) or region. This representation of the apparent motion field only makes it possible to model and identify a constant and purely translational motion parallel to the image plane by object (region, block...) which constitutes a very restrictive class of 3-D motions of an actual natural scene. Let us recall that in the case of sensor motions which are not purely translational parallel to the image plane, which is often the case in televisual scenes (tilt, panning, translations parallel to the optical axis...), the apparent motion vector field cannot be correctly represented on regions or blocks by a simple 2-D translation.

As far as modelling and identification of 3-D motion parameters are concerned, there are several possibilities. Firstly (Section 8.1), we recall the geometrical relations between 3-D motions, 3-D structures (*i.e.*, 3-D geometry of objects) and apparent 2-D motions in the case of the visual "perspective" projective system. The particular cases of the description of objects by planar facets and low-order parametrized approximation of motion vector fields (1st order: affine models and 2nd order: quadratic models) are more particularly detailed.

As far as the resolution methods and the application frameworks envisaged are concerned, we will present separately:

- the monocular case where a unique sensor (if necessary moving) perceives the dynamic scene and, through spatio-temporal observations, tries to reach both motion information and that concerning the structure of objects. The applications within coding schemes concern compression methods ("second" generation with very low rates) or techniques of analysis/synthesis by extraction of high-level global primitives.

- the stereoscopic case where several sensors (2 or even 3 sensors) simultaneously perceive the same dynamic scene which makes it possible to identify, either in parallel or jointly, the structural and motion parameters of the 3-D objects which constitute the scene. Many studies have been carried out into stereo-motion cooperation within the field of Artificial Vision, primarily with the aim of 3-D reconstitution of objects or of robot navigation in complex environments. More recently, for 3-D TV or stereoscopic sequence dynamic restitution applications (CAD of 3-D objects, computer-assisted chirurgical operation...), these techniques have also been studied with the aim of improving image reconstitution quality after analysis/synthesis phase or compression/decompression. Whilst still remaining at the heart of similar motion estimation schemes, the bi- or tri-nocular stereoscopic case makes it possible to improve the observation space and to solve some ambiguities in temporal occlusion regions.

  Some results of simulation of predictive coding schemes with motion compensation will be given, which enables to measure the performance of associated estimators.

1 Models and descriptors of 3-D motions

1.1 Relations between 3-D motions and apparent motions

Let us recall the geometric relations which link the 3-D motion vector $\vec{V} = (\dot{X}, \dot{Y}, \dot{Z})^t$ of a point $(X, Y, Z)^t$ of the surface of an object in motion and its projection $(\dot{x}, \dot{y})^t = (u, v)^t$ on the image plane.. We examine the case of the perspective projective system where

$$x = f\frac{X}{Z} \; ; \; y = f\frac{Y}{Z} \tag{1}$$

In order to simplify the notations, we will select the term f to designate the ratio focal length/pixel size as having a normalized value of 1.

The 3-D motion vector $\vec{V}$ can be expressed using the instantaneous translation vector $\vec{T}$ and of the instantaneous rotation vector $\vec{\Omega}$ of the kinematic screw associated with the moving object [26], *i.e.*,

$$\vec{V} = \vec{T} + \vec{\Omega} \wedge \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{2}$$

which is expressed, by components, as

$$\begin{bmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{bmatrix} = \begin{bmatrix} T_X \\ T_Y \\ T_Z \end{bmatrix} + \begin{bmatrix} \Omega_Y Z - \Omega_Z Y \\ \Omega_Z X - \Omega_X Z \\ \Omega_X Y - \Omega_Y X \end{bmatrix} \tag{3}$$

In the same way, the components of the apparent motion vector associated with the point (x, y) in the image plane, are defined in the case of perspective projection by

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \frac{\dot{X} - x\dot{Z}}{Z} \\ \frac{\dot{Y} - y\dot{Z}}{Z} \end{bmatrix} \tag{4}$$

which after replacement in Equation (4) of the expressions defined in Equation (3) gives

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \frac{T_X}{Z} + \Omega_Y - \frac{T_Z}{Z}x - \Omega_Z y - \Omega_X xy + \Omega_Y x^2 \\ \frac{T_Y}{Z} + \Omega_X - \frac{T_Z}{Z}y + \Omega_Z x + \Omega_Y xy - \Omega_X y^2 \end{bmatrix} \tag{5}$$

The relations (5) are fully specified when the term $1/Z$ is also expressed as a function of the local pixel-coordinates (x, y). In order to retain a maximum quadratic order in Equation (5) as a function of the coordinates (x, y), but particularly since the structural terms of a geometric surface greater than order 1 are difficult to identify without bias on real images, *a priori* hypotheses concerning the regularity of surfaces are given. Then, if the term Z (and, therefore, the term $1/Z$) is expressed by a first order Taylor development,

$$\begin{aligned} Z &= Z_0 + (\tfrac{\partial Z}{\partial X})_0 X + (\tfrac{\partial Z}{\partial Y})_0 Y + o^2(X, Y) \\ &= Z_0 + Z_1 X + Z_2 Y + o^2(X, Y) \end{aligned} \tag{6}$$

it leads to

$$\frac{1}{Z} = \frac{1}{Z_0}(1 - Z_1 x - Z_2 y) + o^2(x, y) \tag{7}$$

subsequently noted by

$$\frac{1}{Z} = n_X x + n_Y y + n_Z + o^2(x, y) \tag{8}$$

(n_X, n_Y, n_Z) specifying the terms of the structure of the local surface which is approximated here by a planar facet (Equation (6)) around (X_0, Y_0, Z_0). Currently, the reference point selected will be the center of gravity of the region for which planar facet approximation (6) or (8) is carried out, being

$$(x_g, y_g)^t = \left(\frac{X_0}{Z_0}, \frac{Y_0}{Z_0}\right)^t \tag{9}$$

1.2 Affine and quadratic models

Equation (5) linking the apparent motion components $(\dot{x}, \dot{y})^t$ to the pixel coordinates and the surface approximation carried out in (8) making it possible to establish a quadratic relation between $\vec{v} = (\dot{x}, \dot{y})^t$ and the coordinates of the point where this measurement is carried out

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} a_1 + a_2 x + a_3 y + a_7 xy + a_8 x^2 \\ a_4 + a_5 x + a_6 y + a_8 xy + a_7 y^2 \end{bmatrix} \tag{10}$$

where,

$$\begin{cases} a_1 = T_X n_Z + \Omega_Y \\ a_2 = T_X n_X - T_Z n_Z \\ a_3 = T_X n_Y - \Omega_Z \\ a_4 = T_Y n_Z - \Omega_X \\ a_5 = T_Y n_X + \Omega_Z \\ a_6 = T_Y n_Y - T_Z n_Z \\ a_7 = -T_Z n_Y - \Omega_X \\ a_8 = -T_Z n_X + \Omega_Y \end{cases} \tag{11}$$

1.2.1 Justification of the linear approximation

Two sub-models of the motion vector local field can be introduced naturally from Equation (10).

1. a linear model: (dim=6) restricts itself to motion parameters $(a_1, a_2, a_3, a_4, a_5, a_6)$.

   This model is also called an affine model in so far as it makes it possible to identify an affine pixel-based transformation. In fact, if the pixel $p_{t+\Delta t} = (x_{t+\Delta t}, y_{t+\Delta t})^t$ is matched to the pixel $p_t = (x_t, y_t)^t$ by the affine relation

   $$p_{t+\Delta t} = \mathrm{A} p_t + \mathrm{B} \tag{12}$$

   $$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} \simeq \frac{1}{\Delta t}(p_{t+\Delta t} - p_t) = \frac{1}{\Delta t}(\mathrm{A} - \mathbb{I}_2) \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \frac{\mathrm{B}}{\Delta t} \tag{13}$$

   we again find the linear relation between motion vector field and pixel-coordinates. An important consequence of this observation is that, when such a linear motion model is used, the properties of affine transformations will be used, implicitly: in particular, let us refer to the transformation of a linear segment in a linear segment, of a polygonal region in a polygonal region and the maintenance of convexity property.

2. a quadratic model (dim=8) using all the parameters $\{a_i\}_{i=1,\ldots,8}$ defined in Equation (10).

   We will then see that these models, even if they prove to be more complete, come up against two major problems: it turns out that it appears difficult to obtain an accurate estimation of quadratic terms from previously estimated 2-D apparent motion measurements; the model described by the Equation (10) is already a restrictive model compared to a general quadratic model which would contain six quadratic terms and is only obtained by first order approximation of local surfaces and rigid motion hypothesis; secondly, the use of a quadratic parametric model in motion compensation only brings minor improvements in the regions of complex motions and can even prove to be less efficient than the use of a lower order parametric model.

1.2.2 Illustration of particular cases of linear modelling

Case 1: If the instantaneous rotation vector $\vec{\Omega} = (\Omega_X, \Omega_Y, \Omega_Z)^t$ is equal to $(0, 0, \Omega_Z)^t$, that is to say where only rotations around the center of gravity of the region, and with a rotational axis parallel to the optical axis are authorized, then the development (10) becomes:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} T_{X_g} \\ T_{Y_g} \end{bmatrix} + \begin{bmatrix} k & -\theta \\ \theta & k \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} + \begin{bmatrix} T_X n_X & T_X n_Y \\ T_Y n_X & T_Y n_Y \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} \tag{14}$$

with

- $(T_{X_g}, T_{Y_g})^t = (a_1, a_4)^t = (T_X n_Z, T_Y n_Z)^t$, translation vector of the center of gravity of the region which, as we note, in relation to the 3-D translation components, is only defined to within one n_Z factor (similarity factor on the Z axis).
- $k = -T_Z n_Z$ and $\theta = \Omega_Z$, terms which are very often preponderant in translation and rotation along the optical axis
- the other terms constituting crossed motion and structure terms along the other axes

Case 2: Simplified linear model (SLM model)

An even rougher form of modelling of the structural geometry of objects and regions consists of considering the scene as a succession of planar facets parallel to the image plane, in the same way as a z-buffer in infography. This leads to $n_X = n_Y = 0$ and, consequently,

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} T_{X_g} \\ T_{Y_g} \end{bmatrix} + \begin{bmatrix} k & -\theta \\ \theta & k \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} \tag{15}$$

The merit of this form of modelling is that it provides a compact representation (4 parameters) for the description of the field and a simple interpretation concerning the 3-D motion components: T_X, T_Y, T_Z, and $\Omega_Z = \theta$.

Case 3 : Constant Model (CST model)

Finally, let us recall the case of the constant model, restriction of the linear model solely to 0 order terms. This model, which is widely used in motion compensation by regions nevertheless proves limited in identifying complex global 3-D motions.

1.3 Linear approximation of the motion vector field and choice of $2\frac{1}{2}$-D descriptors

The analysis base for specifying the geometry of the motion vector field as specified by the Equation (10) is not of course unique. To convince ourselves of this it is possible, through differential operators, to return to the general formulation of a vector field with, for example, linear geometry

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_g \\ v_g \end{bmatrix} + \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} \tag{16}$$

which corresponds to a development limited to first order of the field around the point (x_g, y_g), or

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_g \\ v_g \end{bmatrix} + \mathrm{M} \times \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} \tag{17}$$

François et Bouthemy [13], and Simard and Mailloux [44] recall that the M matrix can be re-written as:

$$\begin{aligned} \mathrm{M} &= \tfrac{1}{2} trace(\mathrm{M})\mathbf{I} + \tfrac{1}{2}(\mathrm{M} - \mathrm{M}^T) + \tfrac{1}{2}(\mathrm{M} + \mathrm{M}^T - trace(\mathrm{M})\mathbf{I}) \\ &= \tfrac{1}{2} div \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \tfrac{1}{2} rot \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + \tfrac{1}{2} hyp_1 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \\ &\quad + \tfrac{1}{2} hyp_2 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{aligned} \tag{18}$$

which makes it possible to introduce general differential operators for the description of a vector field (not necessarily linear) at each (x, y) point

$$\begin{array}{rrcl} divergence = & div(u,v) & = & \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \\ rotational = & rot(u,v) & = & \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \\ hyperbolic\ 1 = & hyp_1(u,v) & = & -\frac{\partial v}{\partial y} + \frac{\partial u}{\partial x} \\ hyperbolic\ 2 = & hyp_2(u,v) & = & \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \end{array} \tag{19}$$

Examples of synthetic fields are provided by Figure 1 and illustrate fairly well the physically interpretable nature of these differential descriptors.

Using these, we thus specify a linear geometry motion vector field by

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} u_g \\ v_g \end{bmatrix} + \frac{1}{2} \begin{bmatrix} (div + hyp_1) & (hyp_2 - rot) \\ (rot + hyp_2) & (div - hyp_1) \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} \tag{20}$$

The analogy with the affine decomposition model defined at Equation (10) makes it possible to define the change of basis between descriptor sets.

$$\begin{cases} a_1 = u_g = T_{x_g} \\ a_2 = (\frac{div + hyp_1}{2}) \\ a_3 = (\frac{hyp_2 - rot}{2}) \\ a_4 = v_g = T_{y_g} \\ a_5 = (\frac{rot + hyp_2}{2}) \\ a_6 = (\frac{div - hyp_1}{2}) \end{cases} \iff \begin{cases} u_g = a_1 \\ div = a_2 + a_6 \\ rot = a_5 - a_3 \\ v_g = a_4 \\ hyp_1 = a_2 - a_6 \\ hyp_2 = a_3 + a_5 \end{cases} \tag{21}$$

According to the estimation method (evoked in Section 8.2) and the intended application (qualitative interpretation and/or use in motion compensation), it would be advisable to select whichever set of descriptors proves to be the most effective. Finally, let us stress that the particular case of linear models defined by Equation (15) corresponds to the case in which the hyperbolic terms (hyp_1 and hyp_2) are disregarded, that is to say:

$$\begin{cases} a_2 = a_6 = \frac{1}{2}\ div \\ a_3 = -a_5 = -\frac{1}{2}\ rot \end{cases} \tag{22}$$

1.4 Design and use of an apparent motion model hierarchy

Up until now, studies carried out in the field of motion estimation-compensation only used a pre-defined motion model, without seeking to adapt it to the various motions present within the image. Let us note that, as a general rule, it is the region-constant model which is used. Now, as there are generally several different types of motions in a single natural image sequence, it would seem to be interesting to adapt the motion model to be identified locally, this, essentially, for the following two reasons:

- the identification of a too simple motion model (for example a constant model) in a region in which the physically observed motions are complex (some sort of 3-D motion of a rigid body for example) can only lead to poor reconstitution by motion compensation or to an over-segmentation of the region (possibly down to pixel level) costly in terms of volumes of motion information to estimate and to transmit (see Figure 1).

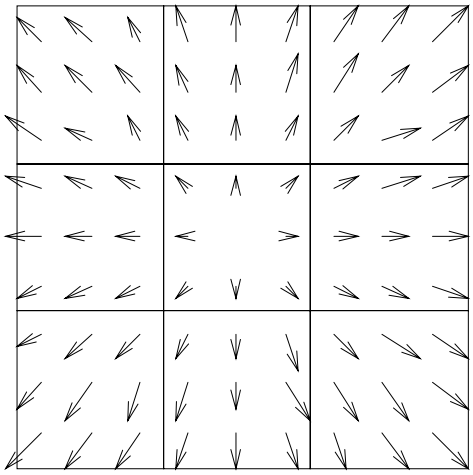

Figure 1: Illustration of the effect of the selection of a model on segmentation: if a divergence model is used, the whole of the vector field constitutes a single homogenous region, on the other hand, if a constant model is used, it is necessary to decompose the main region into several sub-regions (thus more descriptors are used) and that for a less effective result.

- the identification of a sophisticated motion model (for example a quadratic model) on a region in which a single motion can be observed (for example 2-D translation motion parallel to the image plane) leads to large estimation bias, including on the significant parameter sub-vector corresponding to the single motion which naturally should be identified. In fact, as we will establish in the next paragraph, the criterion to be minimized in the motion parameter vector estimation diagram is very often global, since it is simultaneously dependent on all the components of the motion vector to be identified. Thus the components which are not actually observable introduce bias on the identification of the components of the true motion.

Naturally, paragraphs 8.1.2 and 8.1.3 introduced several motion models of increasing complexity. Figure 2 illustrates how these different models can be placed in a hierarchy from the most simple (zero motion) to the most complex. As with [8] and [39], we have included the possibility of introducing into the motion parameters vector to be identified, an estimate of the illumination variation, considered as a potential source of temporal change in the intensity function. Once this model hierarchy has been identified (denoted by $\mathcal{M}_\beta$), it is advisable to define the path strategy within this hierarchy. The introduction of the notion of local adaptivity of motion models signifies the choice from amongst the $\mathcal{M}_\beta$ entity of the most "probable" model β in the sense of a cost or performance criterion for the model β. This cost function very often depends on:

- the error due to reconstitution by motion compensation associated with the model β
- the cost of representation (indeed of transmission if the motion vector field is transmitted in accordance with the coding schemes considered) of the motion information (parameters vector Θ_β with dimensions which vary depending on the model)

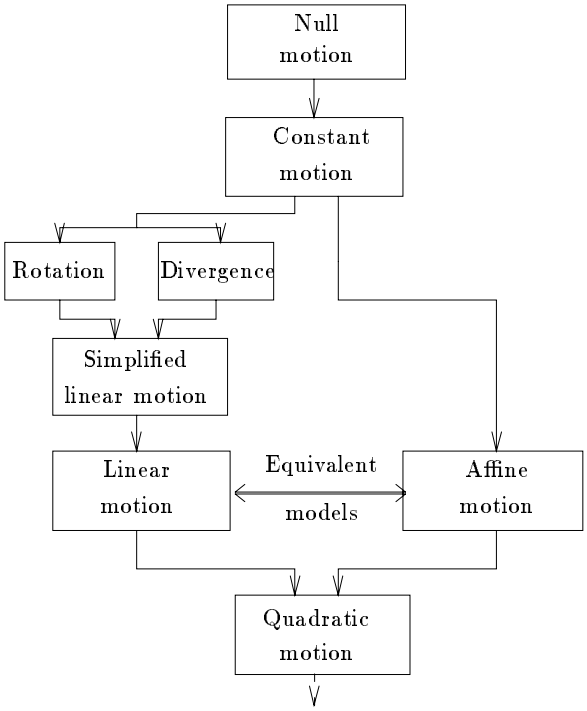

Figure 2: A model hierarchy

- the size of the region considered in order to avoid an under- or over-segmentation of the image
- the operational cost of the identification of the vector Θ_β

It is easy to distinguish two extensive methodologies for the effective use of the $\mathcal{M}_\beta$ entity of motion models:

1. Parallel approach: a test in parallel of all motion models is carried out, region by region in the sense of a MAP criterion, and the most effective model is selected. The clearly formalized mathematical framework of the statistical criteria based to the information theory [40] makes it possible to solve this problem.
2. Sequential approach: this involves using the hierarchy of $\mathcal{M}_\beta$ models in accordance with a pre-defined path which can be either:
   - from the simplest to the most complex model ("coarse to fine" approach)
   - from the most complex to the simplest by progressive suppression of the components of the motion vector ("fine-to-coarse" approach)
   - from an averagely complex intermediate model (for example an SLM model introduced in paragraph 8.1.2) to a more complex or more simple version.

For all these sequential approaches, the mathematical framework for the tests of the hypotheses based on likelihood functions appears well adapted: two hypotheses will be tested by comparison with each other, for example in the sense of maximum likelihood:

- Hypothesis H0: the motion of the current region corresponds to a motion model β

- Hypothesis H1: the motion of this same region corresponds to a just slightly more complex ($\beta + 1$) motion model.

In conclusion, let us note that within the context of the use of such a motion hierarchy, the representation of the motion information will consist of two information fields:

- the map of models selected (one label $\{\beta\}$ per region)
- the motion parameter vector field itself. Let us also recall that the size of the vector Θ_β varies depending on β.

2 Estimation methods in the monocular case

2.1 Estimation of the sensor motion of a static scene

Several motion estimation algorithms try, before or at the same time as the estimation of a dense motion information field, at all points or in all regions of the image, to estimate the sensor motion, in order to be able to identify not the relative motions between the camera and the objects, but the absolute motions of the objects in relation to a fixed reference.

A priori, the camera has freedom of motion throughout the six dimensions of a true motion (3-D translation and 3-D rotation). According to certain hypotheses (see [16], [50], [39]) involving, in particular, the relative distancing of objects present in the scene in relation to the small angles of rotation during a panoramic motion of the sensor, the camera motions can be reduced to the following three classes:

- translations parallel to the image plane (including panning).
- translations perpendicular to the image plane (divergence) analytically equivalent to a change in focal length (zoom).
- rotations around the optical axis.

It can thus be seen that a simplified linear motion model (SLM model with $\Theta_{SLM} = (t_x, t_y, k, \theta)$), as introduced by Equation (15), makes it possible to identify such a sensor motion.

This sensor motion can be estimated directly by one of the methods introduced in the paragraph below. The entire image is then considered as a single region whose center of gravity is the center of the image also identified at the projection of the optical center. Other quantitative information (localization of fixed objects in the scene whose apparent motion is thus not due to the sensor motion alone) or qualitative information (known nature of the sensor motion model) can be injected easily into the algorithm, in order to ease and improve the estimate. *A priori*, such knowledge is rarely available in the case of communication services (contribution, distribution, storage services, etc...) which is the opposite of applications which use "closed-loop" dynamic imagery, that is to say where information is available concerning the sensor motion from its own control (*e.g.*, tele-monitoring, vision for robotics, etc...).

The results in Figures 3 to 7 illustrate the performance obtained when sensor motion is taken into account, in terms of compactness of motion representation and of the error due to reconstitution by motion compensation, in the limited case in which only this sensor motion estimation is carried out.

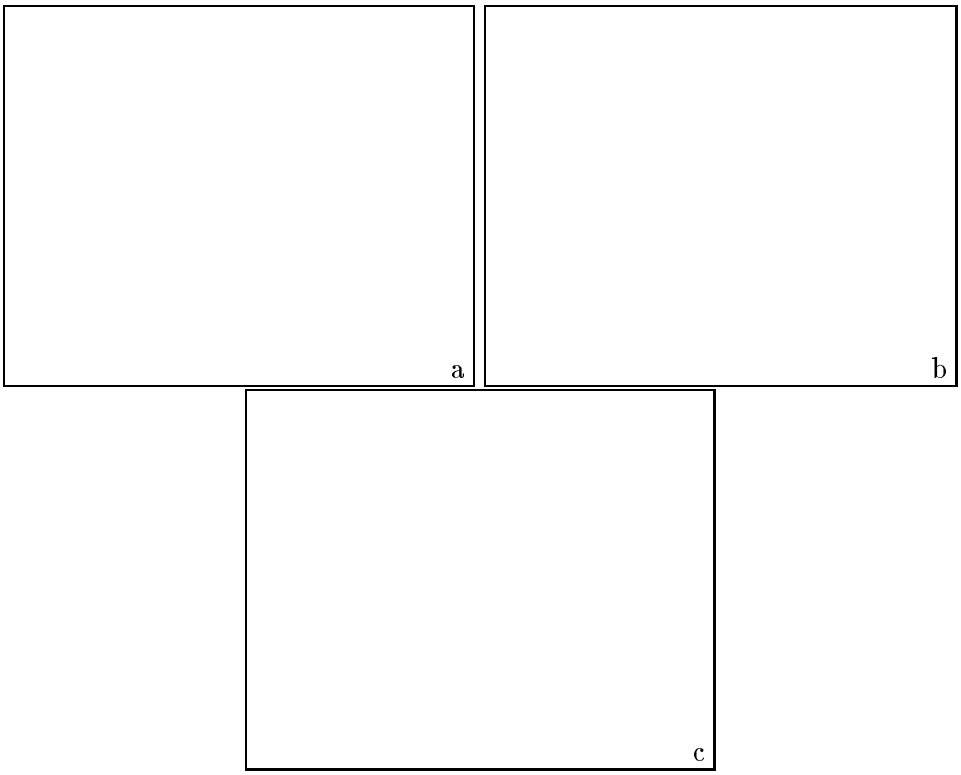

Figure 3: (a) and (b), two original frames of the "Kiel harbour" sequence, (c) Frame difference image with MSE=922.5

2.2 Estimation methods of motion descriptors for a moving scene

All the motion estimation methods - closely related to the aspects of segmentation based on motion in the case of motion estimators by regions - were discussed in Chapter 3, essentially using the 2-D constant translation model (t_x, t_y). Let us also recall that the following general classes of motion estimation were presented:

- translation of a 2-D region (whose "block-matching" algorithm is an example)
- pel-recursive algorithms
- iterative algorithms
- analysis of spatio-temporal frequencies
- parametric models
- segmentation/estimation link

Below we detail how these methods can be extended naturally for more complex parametric motion models (already presented in Section 3.3.2.5). However, two cases present themselves depending on the existence or otherwise of a dense apparent motion vector field preliminary to the estimation of the parameters of more global models. We deal briefly with the case in which such a dense field preexists since, clearly, an algorithmic scheme complete as much for coding as for analysis, will tend to remove itself from the calculation of this dense field, sometimes very operationally complex, if it is not useful. Let us note,

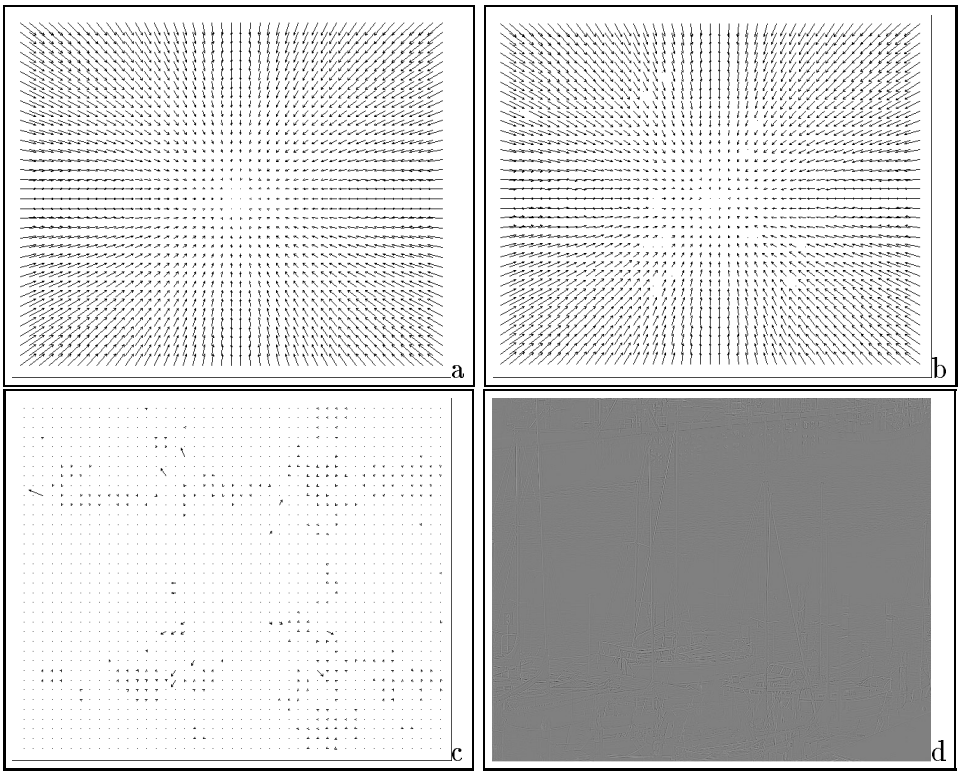

Figure 4: (a) Identification of a global (camera) motion using a divergence motion model, (b) optical flow relative to the global motion, (c) Differential flows, (d) Motion-compensated frame difference image only based to the global motion (a) MSE = 56.3

Figure 5: (a) and (b), two original frames of the "Interview" sequence

however, that through the analytical relations detailed below, it is still possible to pass from a sparse field of motion descriptors to a dense apparent motion vector field and vice versa.

2.2.1 Estimation of a parametric model from a dense motion vector field

As we saw in Chapter 3, many methods make it possible to obtain a dense motion vector field. An illustration is provided below (Figure 8) with the Horn-Schunck algorithm [17]. The idea is to use this dense information in order to extract from it parameters of a more global model (for example an affine or SLM model as illustrated in Figure 9).

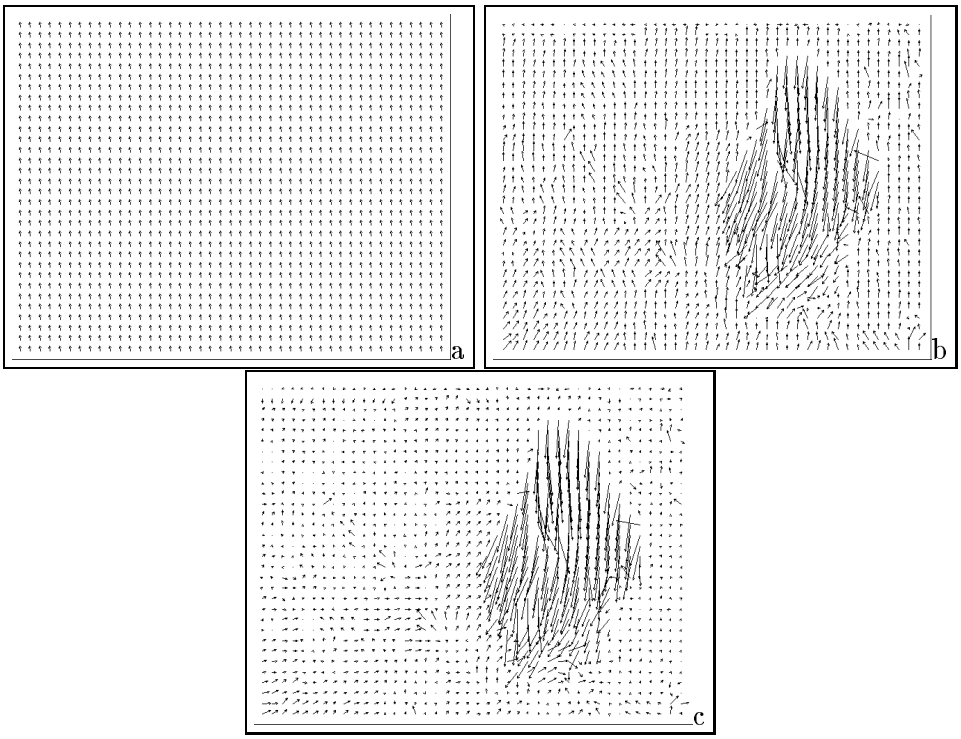

Figure 6: (a) Identification of a global (camera) motion using a constant motion model, (b) optical flow relative to the global motion, (c) Differential flow

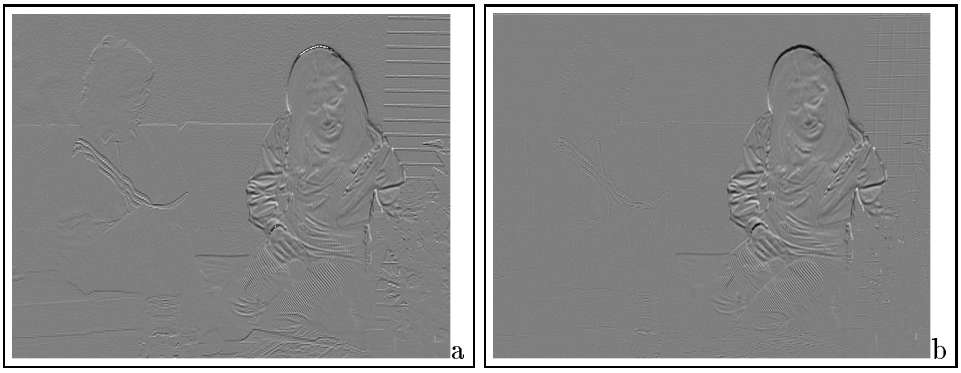

Figure 7: (a) Frame difference image with MSE=137.4, (b) Motion-compensated frame difference image only based to the global motion MSE=100.9

At this stage, we assume that we have image segmentation into homogenous regions in the motion sense. The parameters are obtained:

- by minimization of the mean square error between the initial dense field and the dense field derived from the parametric model ([15], [29], [16]); for example, let us consider an SLM model with parameters $\Theta_{SLM} = (t_x, t_y, k, \theta)^t$ for a region $\mathcal{R}$ and an initial dense field noted as $\{(u_i, v_i)\}$ for each pixel $\in \mathcal{R}$ indexed by i with coordinates (x_i, y_i); the error to be minimized is therefore expressed as:

$$E^2 = \sum_{i \in \mathcal{R}} (t_x + kx_i - \theta y_i - u_i)^2 + (t_y + ky_i + \theta x_i - v_i)^2 \tag{23}$$

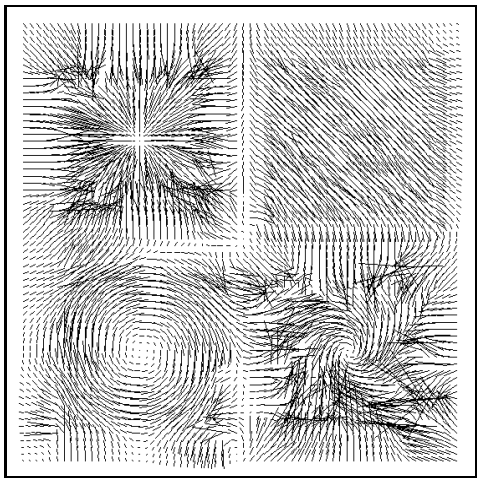

Figure 8: Example of an optical flow obtained by the Horn-Schunck method [17] on differents areas where "pure" divergent, translational, rotational and affine flows have been synthetized

The least mean squares resolution requires the inversion of the 4×4 matrix (for such an SLM model). Simplifications can be made [42] concerning this system's resolution. The resolution equations provide the vector of the following parameters:

$$\left\{ \begin{array}{rcl} t_x & = & \sum_i u_i - k \sum_i x_i + \theta \sum_i y_i \\ t_y & = & \sum_i v_i - k \sum_i y_i - \theta \sum_i x_i \\ k & = & \frac{\sum_i u_i \sum_i x_i - \sum_i u_i x_i + \sum_i v_i \sum_i y_i - \sum_i v_i y_i}{\left(\sum_i x_i\right)^2 - \sum_i x_i^2 + \left(\sum_i y_i\right)^2 - \sum_i y_i^2} \\ \theta & = & \frac{\sum_i u_i y_i - \sum_i u_i \sum_i y_i - \sum_i v_i x_i + \sum_i v_i \sum_i x_i}{\left(\sum_i x_i\right)^2 - \sum_i x_i^2 + \left(\sum_i y_i\right)^2 - \sum_i y_i^2} \end{array} \right. \quad (24)$$

- by separable identification of global translation motions and rotation/divergence in relation to the center of gravity of the region considered, by simple averaging of local estimates [37] the following global parameters are obtained:

$$\left\{ \begin{array}{rcl} t_x & = & \sum_i u_i \\ t_y & = & \sum_i v_i \\ k & = & \sum_i \frac{x'_i(u_i - t_x) + y'_i(v_i - t_y)}{x'^2_i + y'^2_i} \\ \theta & = & \sum_i \frac{x'_i(v_i - t_y) - y'_i(u_i - t_x)}{x'^2_i + y'^2_i} \end{array} \right. \quad (25)$$

where (x'_i, y'_i) represents the relative coordinates in relation to the center of gravity of the region considered.

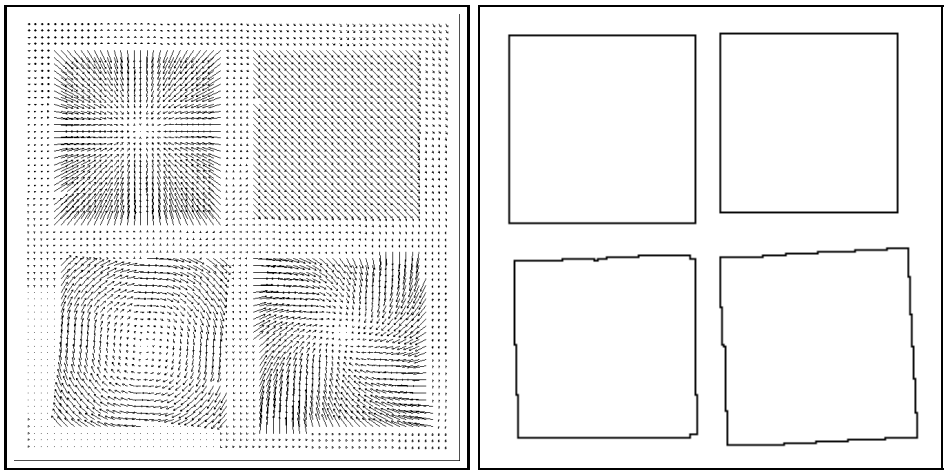

Figure 9: Identification of the affine motion model descriptors on the four regions (Velocity field obtained by using the system in Equation (25))

2.2.2 Direct parametric estimation

Least Mean Square Estimation By extension of the methods introduced in Chapter 3 (paragraph 3.3.2.5) it is quite possible to envisage the introduction into the resolution model of a more complex model (ex. here of an affine model). The resolution of the motion constraint equation is expressed by: for the region $\mathcal{R}$, the optimal estimated motion $\Theta^*_{\mathcal{R}}$ will be

$$\begin{aligned}\Theta^*_{\mathcal{R}} &= ({a_1}^*, {a_2}^*, {a_3}^*, {a_4}^*, {a_5}^*, {a_6}^*)^t \\ &= \arg\min_{\Theta} \sum_{p\in\mathcal{R}} (I_x(p)u(\Theta) + I_y(p)v(\Theta) + I_t(p))^2 \end{aligned} \tag{26}$$

with $u(\Theta) = a_1 + a_2 x + a_3 y$ and $v(\Theta) = a_4 + a_5 x + a_6 y$ (affine model). The least squares resolution is achieved by resolution of a linear system of six equations. Certain simplifications have been proposed [42], [15].

Estimation by a generalized gradient method (see Chapter 3)

Here we seek the solution minimizing the motion compensation mean square error across the whole of the region $\mathcal{R}$ by the gradient optimization technique.

$$\begin{aligned}\hat{\Theta} &= \arg\min_{\Theta} \sum_{p\in\mathcal{R}} DFD^2(p,\Theta) \\ &= \arg\min_{\Theta} \sum_{p(i,j)\in\mathcal{R}} (I(i,j;k) - I(i-u(\Theta), j-v(\Theta); k-1))^2 \end{aligned} \tag{27}$$

The gradient algorithm ([35], [42], [37]) then becomes general to the following iterative estimation process:

$$\vec{\Theta}^{m+1} = \vec{\Theta}^m - \Gamma \frac{\Delta\vec{\Theta}^m}{N_{\mathcal{R}}} \tag{28}$$

$$\text{with } \vec{\Delta\Theta}^m = \sum_{(i,j)\in\mathcal{R}} \begin{bmatrix} \frac{\partial}{\partial a_1} DFD^2(i,j,\vec{\Theta}^m) \\ \vdots \\ \frac{\partial}{\partial a_n} DFD^2(i,j,\vec{\Theta}^m) \end{bmatrix}$$

where,

- m designates the iteration index
- $N_{\mathcal{R}}$ the size of the region $\mathcal{R}$
- Γ a gain matrix which can be either fixed, adaptive, full or empty; limited to a diagonal matrix, the corrective term $\Gamma\vec{\Delta\Theta}^m$ between two iterations is carried out in the direction of the gradient of each component..

In the case of an affine model where $\vec{\Theta} = (a_1, a_2, a_3, a_4, a_5, a_6)^t$, the estimation of $\vec{\Theta}$ is obtained iteratively by:

$$\vec{\Theta}^{m+1} = \vec{\Theta}^m - \sum_{(i,j)\in\mathcal{R}} \Gamma\vec{\phi}^m(i,j) DFD((i,j),\Theta^m) \tag{29}$$

with the displaced gradient vector $\vec{\phi}^m$ equal to

$$\vec{\phi}^m(i,j) = \begin{bmatrix} I_x(i-u(\Theta^m), j-v(\Theta^m); k-1) \\ iI_x(i-u(\Theta^m), j-v(\Theta^m); k-1) \\ jI_x(i-u(\Theta^m), j-v(\Theta^m); k-1) \\ I_y(i-u(\Theta^m), j-v(\Theta^m); k-1) \\ iI_y(i-u(\Theta^m), j-v(\Theta^m); k-1) \\ jI_y(i-u(\Theta^m), j-v(\Theta^m); k-1) \end{bmatrix} \tag{30}$$

The Γ gain matrix is taken diagonal in order to avoid interaction between the different descriptors, otherwise the corrective term $\vec{\Delta\Theta}^m$ would not be taken in the direction of the gradient. Elsewhere, in practice, it is necessary to take account of the difference in scale and in physical size which exists between the various components of the vector $\vec{\Theta}$ of motion parameters. Thus the "constant" parameters (a_1 and a_4) of a affine model will be allocated a larger gain than the other descriptors.

The estimation-segmentation link The identification of the previous motion models requires the definition of a segmentation, either prior to, or concomitant with, the motion estimation phase itself, since this operates on an region $\mathcal{R}$ of matched pixels.

Generally speaking, two approaches can be used:

1. the definition of a segmentation which is either arbitrary (decomposition of the image into blocks) or independent of motion (purely spatial segmentation which has the major inconvenience of constituting an over-segmentation from the motion point of view). This segmentation can be either monogrid, or in relation to a pyramid of information [15], [42], a quadtree splitting [39] or a splitting/merging into regions [6], [14].

   In the case of a pyramidal structure, the elements of this structure inherit motion parameter vectors calculated at a coarser level and a correction to this motion prediction is carried out by parametric estimation as described previously.

   Segmentation into a quadtree allows the progressive decomposition of an image into smaller and smaller regions making it possible firstly to identify the more global

attributes and to lead to identification of local motions (even at pixel level, if the quadtree is complete) at the end of the estimation process. Clearly, a splitting criterion has to be defined; it can be based on the following tests of hypotheses:

- test of a region's homogeneity
  The test consists of comparing the motion homogeneity hypothesis (the $\mathcal{R}_0$ region corresponds to a Θ_0 parametric model) with that of inhomogeneity (presence of several motions). According to Gaussian hypotheses (and zero-mean laws) concerning associated error functions, the search for maximum likelihood leads to testing the following estimated variance:

$$\frac{1}{N_{\mathcal{R}_0}} \sum_{(i,j)\in\mathcal{R}_0} DFD(i,j,\Theta_{\mathcal{R}_0})^2 > \text{or} < \sigma^2 \tag{31}$$

- test of division of a region into L sub-regions
  In this context, the test consists of comparing the following hypotheses:
  Hypothesis H0: the region $\mathcal{R}_0$ corresponds to a unique parametric model.
  Hypothesis H1: the region $\mathcal{R}_0$ could be decomposed into sub-regions, on each one of which a $\Theta_{\mathcal{R}_l}$ parametric model must be identified.
  Bouthemy and Santillana-Rivero [6] test the case in which the region divides up into two sub-regions. According to the same hypotheses as previously mentioned, the likelihood test between the two hypotheses (hypotheses H_0 and H_1 associated with likelihood functions f_0 and f_1) leads to the following test:

$$\log \frac{f(\hat{\theta_1}, \hat{\theta_2})}{f(\hat{\theta_0})} \overset{H_0}{<} \text{or} \overset{H_1}{>} \lambda \tag{32}$$

  and we obtain the following criterion:

$$N_{\mathcal{R}_0} \log \sigma_0^2 - N_{\mathcal{R}_1} \log \sigma_1^2 - N_{\mathcal{R}_2} \log \sigma_2^2 \overset{H_0}{<} \text{or} \overset{H_1}{>} \lambda \tag{33}$$

  where,
  - $N_{\mathcal{R}_0}, N_{\mathcal{R}_1}, N_{\mathcal{R}_2}$ designate respectively the surfaces of the regions $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2$
  - $\sigma_i^2 = \frac{1}{N_{\mathcal{R}_i}} \sum_{p\in\mathcal{R}_i} DFD^2(p, \hat{\Theta}_i)$
    i.e., after linearization,
    $\sigma_i^2 = \frac{1}{N_{\mathcal{R}_i}} \sum_{p\in\mathcal{R}_i} \left(I_x(p)u(\hat{\Theta}_i) + I_y(p)v(\hat{\Theta}_i) + I_t(p)\right)^2$

2. Markovian models make it possible to specify effective observation interaction models (linked with spatio-temporal gradients) and labels (in our case motion parameters). François [14] thus defines a motion based segmentation by Markovian approach using an energy function composed of two terms:

   - one term favouring identical labelling of two adjacent sites (region merging approach)
   - one term seeking to maximize the likelihood of the observations depending on the labels (same formula as previously for σ_i^2)

   A deterministic relaxation scheme makes it possible to propagate labels.

In conclusion, in the case of the use of a motion parametric model in a motion compensation scheme, it seems important :

- to select the criterion to minimize as a direct function of the local compensation errors, *i.e.*, $\sum_{(i,j)\in\mathcal{R}} DFD(i,j,\Theta_{\mathcal{R}})^2$
- to smooth out the motion parameters field to achieve better compactness of presentation.
- to avoid the convergence of the estimation process towards local minima of the non-convex functional to be minimized. The latter two constraints are simply resolved by the introduction of a relaxation algorithm.
- to proceed with a "coarse-to-fine" analysis in a pyramidal or progressive region splitting sense.

Several authors [15], [16], [42], [22], [39] have adopted these principles and obtain interesting results from the point of view of both vector field regularity and motion compensation effectiveness. In Figure 9 we illustrate the example of the algorithm [37], [38] which will serve as a basis for the results on real sequences in paragraph 8.2.4 and Figures 10 and 11.

2.3 Model hierarchy

In the case where, for a given region $\mathcal{R}$, a notion of adaptation of a model to the region is envisaged, it is best to define a selection criterion for the optimum β^* model from all the parametric models written $\mathcal{M}_\beta$. Two families of criteria can be used depending on the sequential or parallel approach desired (see Section 8.1.4).

1. Likelihood ratio

   The procedure is identical to that described previously in the context of splitting of regions. It is a matter of testing, for the same estimation surface (the current region $\mathcal{R}$), two hypotheses:

   Hypothesis H1: the use of a "complex" model
   $\Theta_{\beta_1} = (a_1, \ldots, a_r, \ldots, a_n)^t$

   Hypothesis H0: the use of a "simple" model, restriction to r parameters ($r < n$) of the previous model $\Theta_{\beta_0} = (a_1, \ldots, a_r)^t$

   β^* will be selected in accordance with the most probable hypothesis by comparison with a threshold of the likelihood ratio associated with the two hypotheses.

   According to certain hypotheses (see [14]), it has been shown that this ratio L can be written in the form

   $$L = \frac{N_{\mathcal{R}}}{2} \log(1 + W) = \log \frac{f_1}{f_0} \tag{34}$$

   where f_0 and f_1 are respectively the likelihood functions under hypothese H_0 and H_1 and where W is proportional to a random process according to a Fisher's law which makes it possible, assuming the prior selection of an error probability α (for

example $\alpha = 0.05$) to fix a direction for the test of the hypotheses. In many cases for coding applications, likelihood functions are relative to the motion-compensated mean square errors.

2. Information statistical criteria

   In this context, it is possible to use Akaike and Rissanen's information statistical criteria [40] which, for a given model, evaluate both its performance and its complexity.

   Generally speaking, these two criteria are expressed in the form:

   $$\text{AKAIKE criterion:} \quad \mathcal{C} = -2\log f(y/\Theta_\beta) + 2\dim(\Theta_\beta) \tag{35}$$

   $$\text{RISSANEN criterion:} \quad \mathcal{C} = -2\log f(y/\Theta_\beta) + 2\dim(\Theta_\beta) \times \log N_{\mathcal{R}} \tag{36}$$

   where $f(y/\Theta_\beta)$ is the likelihood of y conditional to Θ_β. The first terms of these two criteria constitute the model performance measures (likelihood), whilst the second are penalization terms for complex models.

   A practical implementation, in order to obtain motion compensation using a Θ motion model hierarchy, was tested [39] by using a measurement criterion derived from the Rissanen criterion and compatible with the function $(\frac{1}{N_{\mathcal{R}}} \sum_{(i,j)\in\mathcal{R}} DFD^2)$ to be minimized, already used in the Θ_β vector estimation process. This criterion is expressed by:

   $$\mathcal{C}_\beta = \log \frac{1}{N_{\mathcal{R}}} \sum_{(i,j)\in\mathcal{R}} DFD^2((i,j),\Theta_\beta) + \alpha \frac{r(\beta)}{N_{\mathcal{R}}} \tag{37}$$

   where

   - α is a weighting coefficient (for example, $\alpha = 0.1$).
   - $r(\beta)$, motion model encoding rate, represents the volume of binary information in the entropic sense for example, required to represent and transmit the $\vec{\Theta}_\beta$ parameters vector.

   If this criterion is applied to the two motion models 1 and 2, then the model 1 will be selected, if

   $$\mathcal{C}_{\beta_1} > \mathcal{C}_{\beta_2} \tag{38}$$

2.4 Estimation of 3-D motion

The estimation of 3-D motion based on image sequences can be carried out naturally using two distinct approaches. The first of these, called the two-stage method, consists of calculating these 3-D motions from a previously estimated 2-D apparent motion vector field. The second, called the direct method, attempts to evaluate these 3-D motions directly from spatio-temporal derivatives of the intensity function. We describe these two general approaches below.

2.4.1 Two-stage estimation methods

This approach, which is similar to that evoked in paragraph 8.2.2 for the estimation of a $2\frac{1}{2}$-D parametric model from a 2-D motion vector field, is based on the following scheme:

stage 1: estimation of a 2-D displacement vector field which will be sparse (discrete methods of matching 2-D primitives) or dense (differential methods) by one of the estimation methods described in Chapter 3.

stage 2: By equations linking the projected 2-D motions and 3-D motions (see paragraph 8.1.1 in the case of a dense field), this second stage identifies the 3-D motion parameters based on the 2-D primitives' field.

We will deal with the case of discrete methods in Section 8.3, since it is very similar to the problem of stereovision-motion cooperation on discrete primitives. Within the context of differential methods, many authors ([1], [45], [55]) pose the problem of the determination of motion and of the 3-D structure from apparent motion in the form of the minimization of a quadratic criterion based on equations concerning 2-D/3-D relations. Even in accordance with the theory of the observation of rigid objects, Equation (5) shows that this problem of optimization is non-linear.

As an example, in the case of differential methods, Adiv [1] breaks this estimation process down into two stages. The first of these consists of segmenting an apparent motion vector field (assumed to have been calculated previously) into regions corresponding to planar facets. The parametric motion modules are thus quadratic models defined by the equations at (11). The estimation technique is based on a generalized Hough transform; from Equation (5), the energy function Φ is defined by

$$\Phi = \sum_{\mathcal{R}} (u - \alpha_\Omega - \alpha_T z)^2 + (v - \beta_\Omega - \beta_T z)^2 \tag{39}$$

with

$$\left\{ \begin{array}{rcl} \alpha_\Omega & = & -xy\Omega_X + \Omega_Y(1+x^2) - y\Omega_Z \\ \beta_\Omega & = & -\Omega_X(1+y^2) + xy\Omega_Y + x\Omega_Z \\ \alpha_T & = & \frac{T_X - xT_Z}{\|T\|} \\ \beta_T & = & \frac{T_Y - yT_Z}{\|T\|} \\ z & = & \frac{\|T\|}{Z} \end{array} \right. \tag{40}$$

which consists of separating the terms which involves the instantaneous translation vector $T = (T_X, T_Y, T_Z)^t$ and the instantaneous rotation vector $\vec{\Omega} = (\Omega_X, \Omega_Y, \Omega_Z)^t$ respectively. Assuming the constancy of the energy function Φ depending on the relative depth variable z ($\frac{\partial \Phi}{\partial z} = 0$), we can deduce the optimum relative depth

$$z^* = \frac{(u - \alpha_\Omega)\alpha_T + (v - \beta_\Omega)\beta_T}{(\alpha_T^2 + \beta_T^2)} \tag{41}$$

which, carrying over to Equation (39)

$$\Phi = \sum_{\mathcal{R}} \frac{((u - \alpha_\Omega)\beta_T + (v - \beta_\Omega)\alpha_T)^2}{(\alpha_N^2 + \beta_N^2)} \tag{42}$$

The unitary vector $\frac{T}{\|T\|}$ can then be parametered in an angular space (ν, ξ) such that:

$$\frac{T_X}{\|T\|} = \sin\nu\cos\xi \ , \ \frac{T_Y}{\|T\|} = \sin\nu\sin\xi \ , \ \frac{T_Z}{\|T\|} = \cos\nu \tag{43}$$

and the energy function Φ is then parametered to $\Phi(\Omega, \nu, \xi)$.

The generalized Hough transform makes it possible to calculate the optimum couple (Ω^*, T^*) such that

$$(\Omega^*, T^*) = \underset{\nu, \xi}{\arg\min}\ \Phi(\Omega^*, \nu, \xi) \tag{44}$$

On completion of this first stage, a fusion of adjacent components corresponding to the same parametric transformation is carried out, using least squares criteria. The algorithm continues by iterative sequencing of these motion-structure parameter estimation procedures and that of the grouping together of regions which correspond to a single transformation. Adiv [2] extends his work by raising the ambiguities inherent in the estimation of 3-D motion and of depth; these ambiguities are essentially of two types:

- a single 2-D field can have several 3-D interpretations (non-unicity of representation) [2], [5], [51].
- an estimation bias on the 2-D primitives field induces an estimation bias on the 3-D parameters and often creates phenomena of instability in estimations.

2.4.2 Direct estimation methods

These methods seek to mitigate the drawbacks mentioned previously by direct estimation of parameters linked to motions and 3-D structures without previously estimated apparent motion fields. In this context, we again find extensions of estimation methods known in the 2-D case, such as extensive recursive estimation methods in the case of parametric motion models ([36], [9], [10], [39]) and iterative estimation methods based on the "brightness change equation" or extensive motion constraint equation in the case of 3-D motions and particular 3-D structures (planar, quadratic surfaces, ...) [18], [33].

Dugelay and Pele [9], and Netravali and Salz [36] start off from the following three-stage approach:

- from the Equations (11) defining the relations between apparent motion description parameters $\mathcal{A}$, those of 3-D motion $\mathcal{C} = (\vec{\Omega}, \vec{T})^t$ and those of structure $\mathcal{K} = (\frac{n_X}{n_Z}, \frac{n_Y}{n_Z}, 1)^t$
- from an initial vector or previous estimate: $\mathcal{C}^{n-1}, \mathcal{K}^{n-1}$

it is possible to repeat the following three stages:

Stage 1: calculation of $\mathcal{A}^{n-1}$ from $\mathcal{C}^{n-1}, \mathcal{K}^{n-1}$, initial values using Equation (11);

Stage 2: a differential method of estimating a corrective term $\delta\mathcal{A}^{n-1}$ is operated by gradient algorithm as follows (see Equations (28) to (30)):

$$\delta\mathcal{A}^{n-1} = \epsilon \sum_{p \in \mathcal{R}} DFD(p, \mathcal{A}^{n-1}; t-1) \frac{\delta}{\delta \mathcal{A}} DFD(p, \mathcal{A}^{n-1}; t-1) \tag{45}$$

Stage 3: based on the system of Equations (11) calculation of the parameters $\mathcal{C}^n$ and $\mathcal{K}^n$, function of $(\mathcal{A}^{n-1} + \delta\mathcal{A}^{n-1})$.

This system of 8 unknowns and 8 non-linear equations works out by successive linearization (Newton method) for example.

The second family of approaches ([18], [32], [33]) consists of starting with the theory of the temporal invariance of the intensity function expressed by the motion constraint equation

$$I_x u + I_y v + I_t = \vec{\nabla} I . \vec{v} + I_t = 0 \tag{46}$$

In a vectorial manner, Equation (4) deduced from the perspective projection model can be expressed:

$$\vec{v} = \begin{bmatrix} u \\ v \\ 0 \end{bmatrix} = \frac{\vec{z} \wedge (\vec{V} \wedge \vec{P})}{(\vec{P}.\vec{z})^2} \tag{47}$$

where $\vec{p} = (x, y, 1)^t$, $\vec{P} = (X, Y, Z)^t$, $\vec{V} = (\dot{X}, \dot{Y}, \dot{Z})^t$, $\vec{v} = (\dot{x}, \dot{y}, 0)^t$ and $\vec{z}$ is the unitary vector along the optical axis, with $\vec{p} = \frac{\vec{P}}{\vec{P}.\vec{z}}$. By substituting the expression of $\vec{V}$ (Equation (2)) in Equation (47), that gives us:

$$\vec{v} = \vec{z} \wedge (\vec{p} \wedge (\vec{p} \wedge \vec{\Omega} + \frac{\vec{T}}{\vec{P}.\vec{z}})) \tag{48}$$

The motion constraint equation (46) expanded to the 3-D case is then expressed:

$$\vec{\nabla} I . (\vec{z} \wedge (\vec{p} \wedge (\vec{p} \wedge \vec{\Omega} + \frac{\vec{T}}{\vec{P}.\vec{z}}))) + I_t = 0 \tag{49}$$

or in a more compact fashion, if $\vec{s} = (\vec{\nabla} I \wedge \vec{z}) \wedge \vec{p}$ and $\vec{w} = \vec{s} \wedge \vec{p}$, then Equation (49) becomes

$$\frac{\vec{s}.\vec{T}}{\vec{P}.\vec{z}} + \vec{w} \wedge \vec{\Omega} + I_t = 0 \tag{50}$$

The resolution method often assumes a geometric structure model. For example, in the planar region case, we have the region of the 3-D points $\left\{\vec{P}\right\}$ defined by $\vec{P}.\vec{N} = 1$ which is equivalent to $\vec{p}.\vec{N} = \frac{1}{\vec{P}.\vec{z}}$. The motion constraint equation then becomes

$$(\vec{s} \,.\, \vec{T}) \, (\vec{p} \,.\, \vec{N}) + \vec{w} \wedge \vec{\Omega} + I_t = 0 \tag{51}$$

and the resolution into $(\vec{T}, \vec{\Omega}, \vec{N})$ is carried out by iterative resolution of a functional minimization algorithm

$$\mathcal{J} = \int \int_{\mathcal{D}} ((\vec{s}.\vec{T})(\vec{p}.\vec{N}) + \vec{w} \wedge \vec{\Omega} + I_t)^2 dx dy \tag{52}$$

These approaches are thus a direct extension of the iterative estimation method normally used in the 2-D case. Other region models have also been tried [33] such as quadratic patches, cylindrical surfaces, etc.

2.5 Use of motion compensation in a predictive coding scheme

The use of parametric motion models within a predictive coding scheme with motion compensation (see Chapter 4 for an introductory description of these schemes) appears to be a natural extension of the usual case where a dense motion vector field compensates the image. As a matter of fact, as illustrated by Equation (10) in the context of a general quadratic model, if, for each region $\mathcal{R}_m$ of the image, we have the motion parameter vector $(\Theta_{\beta,m})$ identified corresponding to the motion model β, it is always possible to derive a dense apparent motion vector field from the $\{\Theta_{\beta,m}\}$ and use it in a motion-compensated loop.

The prediction by motion compensation will be equal to

$$\hat{I}(i,j,k) = \tilde{I}(i - \hat{u}(\Theta_{\beta,m}), j - \hat{v}(\Theta_{\beta,m}); k-1) \tag{53}$$

for each pixel with coordinates $(i,j)^t$ and where,

- $\tilde{I}$ indicates the previously reconstituted image
- $\hat{I}$ indicates the current image to be predicted
- $(\hat{u},\hat{v})$ the dense field predicted from the field $\{(u,v)\}$ derived from the parameters $\{\Theta_{\beta,m}\}$.

Because of the compact nature of the representation of the motion information which represents the $\{\Theta_{\beta,m}\}$, this information is usually transmitted and in this case $\{(\hat{u},\hat{v})\}$ is selected as being the estimated field: $\hat{u}(i,j) = u(\Theta_{\beta,m})$ and $\hat{v}(i,j) = v(\Theta_{\beta,m})$ for each pixel $(i,j) \in \mathcal{R}_m$.

Let us recall that in such a scheme, the information transmitted has to be decomposed into four parts

1. the image segmentation into N regions $\{\mathcal{R}_m\}_{m=1,\ldots,N}$;
2. the type of model used β_m for each region $\mathcal{R}_m$;
3. the quantized motion parameters vector $(\Theta_{\beta,m})$ for each region $\mathcal{R}_m$;
4. the quantized motion compensation error.

As far as the coding of the segmentation map is concerned, a compromise has to be found between the following two extreme cases:

1. *a priori* known arbitrary segmentation such as a block decomposition: the coding cost for such a segmentation is null;
2. adapted spatial segmentation on all images: consequence of extensive coding due to the fact of the irregularity of the edges obtained.

Binary coding schemes adapted to edges (for example Freeman codes) can be used, even if it could use a lot of bit rate to encode this map of contours. Quadtree decomposition allows good adaptation of the segmentation to the local contents of the image at only a small coding cost expressed by [24], [39], [41], [43]

$$R_{\text{quadtree}} = \frac{4}{3}NR - \frac{1}{3}NR_{init} - NR_{min} \tag{54}$$

if NR, NR_{init}, NR_{min} designate respectively the number of regions within the final image after the quadtree decomposition, the number of regions within the initial image (initial grid) and the number of regions with the minimal size (quadtree roots).

The coding cost of the label β_m designating the motion model selected for the current region $\mathcal{R}_m$ clearly only exists in the case of the use of a distinct motion model hierarchy and can be accessed by an entropy cost.

The parameter vector $\Theta_{\beta,m}$ is transmitted after quantization. Note that the various components of this vector do not require the same accuracy of quantization. Adapted quantizers must be designed for each component.

Finally, the coding of the prediction error by motion compensation uses all coding-source techniques (transform coding, entropy coding, ...) again making it possible to decorrelate the information from a spatial or frequency point of view, and thus to reduce by as much, the transmission cost of this information field. Figure 10 shows , applied to the so-called image sequence *Interview*, motion compensated error image when a motion-based quadtree segmentation is used. Moreover, the distorsion v.s rate trade-off is assessed in Figure 11 for several linear scalar quantization versions of the motion compensated errors.

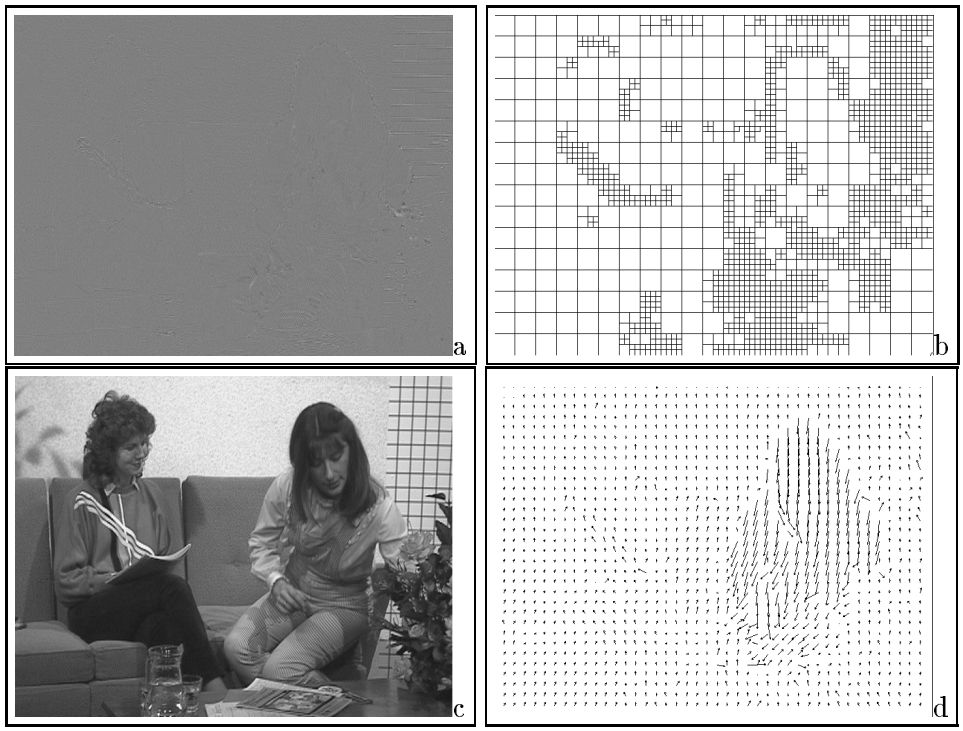

Figure 10: Motion compensation of the "Interview" sequence using a "constant motion" model. (a) Motion-compensated differences : MSE=17.9, (b) quadtree segmentation (4×4 regions are not illustrated), (c) Reconstructed image, (d) Motion vector field

2.6 Use of an analysis-synthesis coding approach

The estimation schemes previously described lend themselves well to the definition of schemes involving object-oriented coding by analysis-synthesis. The first work carried out in this field ([3], [11], [12], [20]) assumed an extensive knowledge of the nature of the objects manipulated and restricted itself to a particular category of scenes such as the

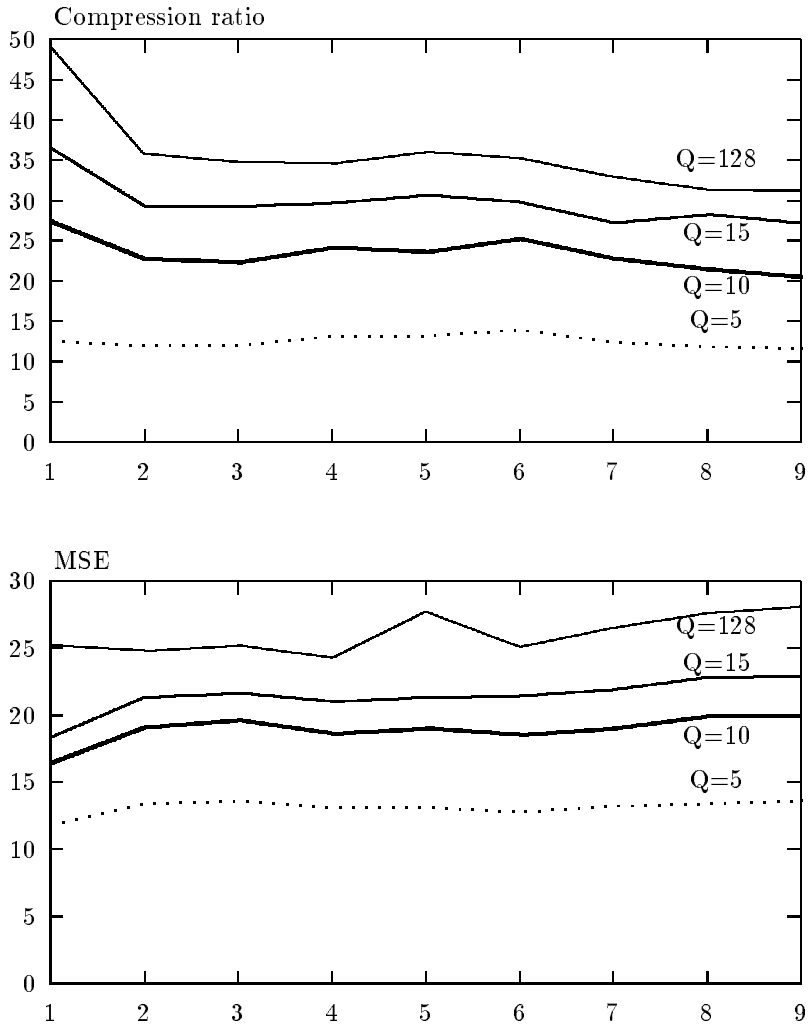

Figure 11: "Interview" sequence. Compression ratio and MSE for different values of the elementary step of quantization for motion-compensated errors.

motion of human faces (videophone services or video conferences with very small rates envisaged). In this case, the hypotheses in the preceding paragraphs, used to establish the relations between 3-D motion/structure and apparent 2-D motion, were valid: rigid objects decomposed into planar surfaces, small rotation angles, small depth variation between two successive images. Musmann *et al* [30] and Hötter [19] develop such an analysis-synthesis object-oriented coding approach, using either the 2-D motion estimation by linear regression methods or the 3-D estimation by prediction/verification methods. The general scheme of the approach is described in the Figure 12. The sequence analysis phase concerns the extraction of three types of information:

- the shape of objects (regions)
- their motion

- the texture or radiosity information

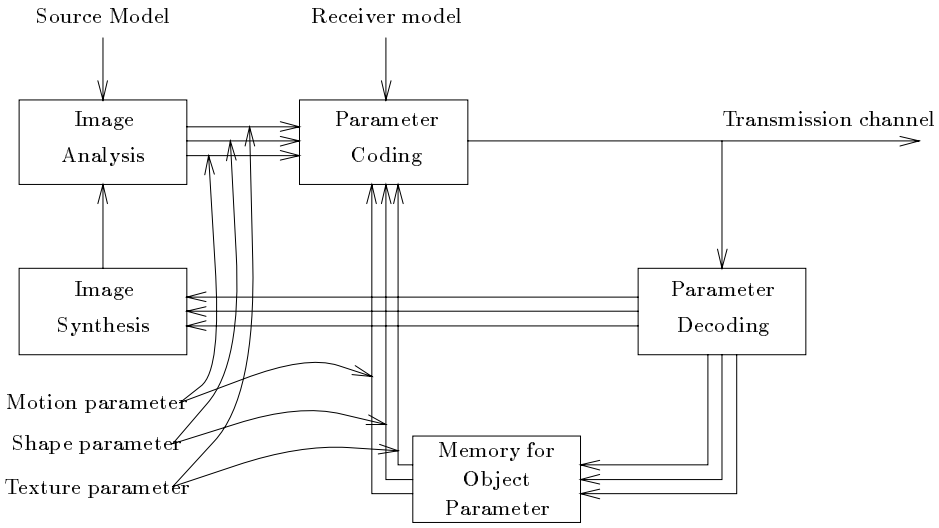

Figure 12: Block-diagram of an object-oriented analysis-synthesis coder

These information fields being different in nature, a specific coding procedure is used for each of them. The shape information describes the outline of objects and this code naturally by contour coding techniques. Only temporal changes in shape will be coded predictively. Motion information also codes predictively in relation to motion parameters estimated on the same object to the previous image. Finally radiosity information can be compressed by hybrid coding techniques with motion compensation.

In conclusion, let us note that these analysis-synthesis coding approaches are often limited to the identification of $2\frac{1}{2}$-D motion parametric models without seeking the whole range of 3-D motion + structure parameters. Such a full range would make it possible to synthesize the scene not only from the true viewing angle at the current moment, but also from all sensor-object relative intermediate positions, which would make it possible to obtain efficient temporal or spatial interpolation schemes. This remains difficult to achieve, however, given the current levels of accuracy obtained on 3-D structural parameters after identification and given that these parameters are only known to a relative depth factor. The stereovision-motion cooperation techniques dealt with in the next section can make it possible in part to overcome these disadvantages.

3 Motion estimation methods in the binocular case

3.1 Introduction

Unlike the monocular case, here we assume the availability of several stereoscopic sensors makes it possible to perceive at different moments (stereocopic sequences) the scene composed of 3-D objects provided with 3-D motions from several points of view. Various experimental contexts can be studied:

- number of sensors: at least two cameras, in order to allow the creation of a stereoscopic effect. This number can be greater (the case of trinocular vision for example

was explored) in order to facilitate the matching phase and to identify certain ambiguities more easily.

- geometry of the stereoscopic system: Most studies which have dealt with this algorithmic theme of stereo-motion cooperation use a stereoscopic system, in which cameras are set out in parallel in a unique plane (*i.e.*, image planes are identical) which assumes a depth focalization at infinity, and where the separation of the geometric base of sensors is large (*i.e.*, greater than the distance corresponding to the visual system of about 65mm). These choices clearly are uncompatible with the optimal conditions of the quality of relief perception (see paragraph concerning the use of these techniques in 3-D TV) for which a respect of different levels of conformity is conventionally introduced.

- calibration of the stereoscopic system: this procedure signifies the prior identification of the intrinsic parameters of each sensor (focal length, coordinates of the optical sensor, radial distortion factor,... see Chapter 1), as well as the extrinsic parameters matching by a geometric screw $(\mathrm{R}_l^r, \mathrm{T}_l^r)$ (3-D rotation + 3-D translation) the relative references attached to each sensor (l = "left" sensor, r = "right" sensor in this paragraph).

  This calibration phase enables:

  - the establishment equations linking the 2-D pixel coordinates to the 3-D point coordinates

    $$\begin{bmatrix} Zx \\ Zy \\ Z \end{bmatrix} = \begin{bmatrix} F_x & 0 & x_c \\ 0 & F_y & y_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{55}$$

    where (X, Y, Z) designate the coordinates of a 3-D point, (x, y) designate the 2-D pixel coordinates and (x_c, y_c, F_x, F_y) are the intrinsic parameters of the sensor (case of a perspective projection sensor model without radial distortion)
  - the passage of "left" coordinate references to "right" and vice versa

    $$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_r = \mathrm{R}_l^r \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_l + \mathrm{T}_l^r \tag{56}$$

  - the definition of epipoles: the right (left resp.) epipole is the projection on the right (left resp.) image plane of the optical center of the left (right resp.) camera. Epipolar lines linking epipoles and optical centers are associated. This epipolar geometry makes it possible to constrain analytically the geometry of the search window during the matching of primitives between left and right images.

  It is clear that in the absence of any calibration, only fairly rough heuristics can be used:

  - selection of the optical center at the center of the image
  - focal parameters fixed without identification
  - search window limited in number of pixels directly in the image plane and hypotheses of horizontal epipolar lines

These heuristic selections naturally introduce large sources of error in the motion estimation and disparity algorithms then used. Tamtaoui [47] carried out a study into the robustness of these algorithms faced with such errors or inaccuracies on the calibration parameters.

Once these experimental selections have been made, the problem of 3-D or 2-D motion estimation in the context of stereoscopic sequences is then posed in these terms: in the short term at two successive moments $(t, t+1)$, as illustrated in Figure 13, we have four observation fields (in the binocular case dealt with here) of a 3-D primitive $\mathcal{P}$ moving in 3-D space, in the case of a rigid object according to the kinematic screw $\vec{V} = (\vec{T}, \vec{\Omega})$; from these four observation fields various 2-D, $2\frac{1}{2}$-D or 3-D information fields can be identified:

- disparity fields ($\{\delta_t\}$ at time t and $\{\delta_{t+1}\}$ at time $t+1$ respectively) by standard matching primitives techniques
- 2-D apparent motion vectors ($\{\Theta_l\} = \left\{\vec{d_l}\right\}$ on the left sequence and $\{\Theta_r\} = \left\{\vec{d_r}\right\}$ on the right sequence respectively) by use of a monocular 2-D apparent motion estimation algorithm
- motion descriptor fields (resp. $\{\Theta_l\}_\beta$ and $\{\Theta_r\}_\beta$) dependent on a previously defined β motion model
- 3-D motion and structure parameter fields in the monocular case applied here to each stereoscopic sequence.

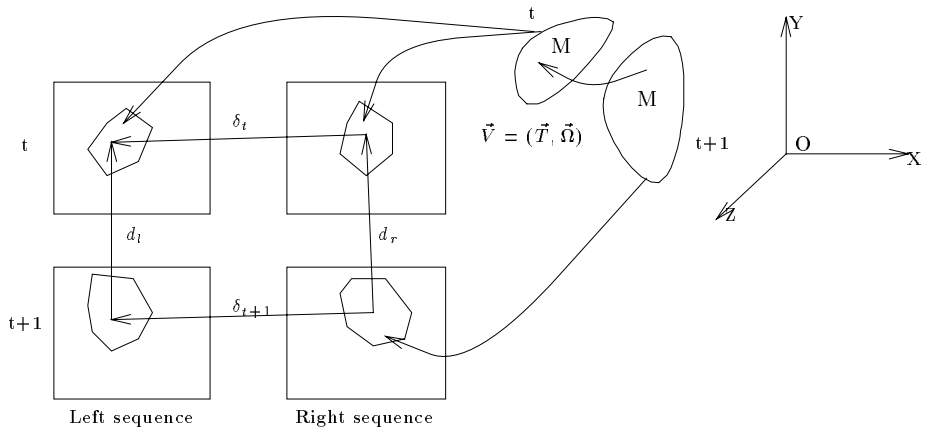

Figure 13: Stereo-motion observation space and associated identifiable information fields

We will not go back over the estimation techniques of these various information fields which have already been studied in Chapter 3 and at the beginning of this chapter in the monocular case. However, let us remember that the manipulated primitives can be of different levels:

- pixel primitives: the information fields are dense
- contour or region primitives: the information fields are sparse.

Below we discuss more particularly the various sequencing or matching possibilities of these stereo-motion primitive estimation procedures; three approaches are distinguished: the first consists of identifying the 3-D motion of objects by temporal matching of 3-D primitives (the "stereo then 3-D motion" approach); the second consists of starting with 2-D apparent motion fields, independently estimated in each stereoscopic sequence, and then raised again by stereoscopic relation between 3-D motion and structure information fields ("3-D motion then stereo" approach); finally, the third approach, which is meant to be better adapted to the case of the use of these motion estimation techniques in a coding context, carries out the joint estimation of motion descriptor fields ("2-D, $2\frac{1}{2}$-D stereo constrained motion" approach) simultaneously in both stereoscopic sequences, by respecting the constraints due to the intrinsic stereoscopic geometry.

3.2 3-D motion by matching 3-D primitives

This approach can be arranged as follows:

Stage 1: After identification of a disparity field $\{\delta_t\}$ (resp. $\{\delta_{t+1}\}$) throughout the sequence, for every stereoscopic couple of images, a depth map is produced $\{Z_t(x,y)\}$ (resp. $\{Z_{t+1}(x,y)\}$) for every image.

Stage 2: A matching phase for 3-D primitives, obtained by successive depth maps, is used.

Stage 3: Instantaneous depth maps and the matching previously carried out make it possible to deduce the 3-D motions + structure of manipulated primitives.

Several authors have studied this type of approach by trying to minimize the number of 3-D primitives to be matched. Leung and Huang [23], Netravali *et al* [34], and Mitiche and Bouthemy [27] worked on 3-D pixel-based primitives; since theoretically three non-colinear points are enough to determine the 3-D motion of a rigid object, a sparse 3-D point depth map is first formulated by stereo-matching. A temporal matching on one of the stereoscopic sequences then makes it possible to identify the 3-D motion of these points. The raising of certain ambiguities is then effected by the verification on the other stereoscopic sequence, of a matching of projected 3-D points. Kim and Aggarwal [21] base their approach on the joint extraction of depth maps on contour-primitives extracted by zero crossings of Laplacians and on pixel-based primitives by Moravec operator. A two-pass relaxation method (in order to ensure the symmetry of temporal matching) is used to link the 3-D primitive maps of two successive images (t) and $(t+1)$; the cost function for the relaxation procedure is based on the notion of motion invariants for rigid bodies such as distance ratios or angles between primitives. Lingxiao *et al* [25] present a method in which the estimation phases of the instantaneous rotation vector and that of translation are uncoupled. Firstly, the centroids of the pixel sets of the left and right views are superposed; on this new set of translated points, the rotation vector $\vec{\Omega}$ calculation is carried out by least mean squares method in the case of a planar structure; finally the translation vector $\vec{T}$ is deduced from Equation (2) itself.

Many other studies have introduced alternative algorithms to those described here. Due to the sparse nature of the processed primitive fields, these stereo-motion cooperation algorithms are intended more particularly for the reconstitution of 3-D objects or as navigation aids for robots by dynamic stereoscopic vision [31], [49]. In stereoscopic

sequence coding, it is still necessary to segment and interpret, in terms of motion and 3-D structures, a complete partition of the images, which makes the two complementary approaches developed below more attractive.

3.3 3-D motion based on 2-D motion fields

Another approach to the calculation of the 3-D motion and structure parameters is based on the independent and prior combination of estimated 2-D apparent motion fields on each of the stereoscopic sequences.

Mitiche [28] starts from the hypothesis of the observation of at least four 3-D points in two stereoscopic sequences. Each point checks the equations

$$\begin{cases} \begin{bmatrix} x_r & y_r & 1 \end{bmatrix} \mathrm{A} \begin{bmatrix} x_l \\ y_l \\ 1 \end{bmatrix} = 0 \\ \begin{bmatrix} u_r & v_r & 0 \end{bmatrix} \mathrm{A} \begin{bmatrix} x_l \\ y_l \\ 1 \end{bmatrix} + \begin{bmatrix} x_r & y_r & 1 \end{bmatrix} \mathrm{A} \begin{bmatrix} u_l \\ v_l \\ 0 \end{bmatrix} = 0 \end{cases} \tag{57}$$

where A, a 3×3 matrix, depends only on the relative displacement $\mathrm{R}_l^r, \mathrm{T}_l^r$ between the systems of coordinates linked to the stereoscopic cameras.

The identification of A (which represents 8 unknown variables after normalization) can be carried out by resolution of the linear system on four observed points. By using the apparent motion field itself, this solves the problem of calibrating the stereoscopic system. For all other 2-D matched point sets, it will therefore be possible to return to the depth information by simple triangulation and thus to obtain access to the 3-D kinematic screw $(\vec{T}, \vec{\Omega})$ by resolution of the system of Equation (5) (linear in $\vec{T}$ and $\vec{\Omega}$) once this depth map is known. Waxman *et al* [53], [54] studied, in particular, the relations between 2-D motion fields. They define the relative flow or binocular difference flow by

$$\Delta\vec{d}(x_l, y_l, \delta) = \vec{d_r}(x_l + \delta(x_l, y_l), y_l) - \vec{d_l}(x_l, y_l) \tag{58}$$

where $\delta(x_l, y_l)$ designates the disparity measure obtained at the current point (x_l, y_l) of the left view; in the case of parallel and aligned cameras (*i.e.*, $Z_l = Z_r$ and $y_l = y_r$ at all points), it is expressed by

$$\delta(x_l, y_l) = \frac{b}{z_l(x_l, y_l)} \tag{59}$$

where b measures the distance (baseline) between the two stereoscopic sensors.

Equation (5) is reformulated, by separating the terms linked with the instantaneous translation $\vec{T}$ and those linked with the rotation $\vec{\Omega}$ by:

$$\vec{d}(x, y) = \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{Z(x, y)} A(x, y).\vec{T} + B(x, y).\vec{\Omega} \tag{60}$$

From Equations (58) to (60), we deduce the following analytical relation between disparity fields, relative flow components and 3-D motion (in the case of aligned cameras):

$$\begin{cases} \frac{\Delta u(x_l, y_l, \delta)}{\delta(x_l, y_l)} = \frac{1}{b} T_Z \delta + y_l \omega_X - x_l \omega_Y \\ \frac{\Delta v(x_l, y_l, \delta)}{\delta(x_l, y_l)} = 0 \end{cases} \tag{61}$$

If a planar structure hypothesis is used, *i.e.*, $\frac{1}{Z(x,y)} = n_X x + n_Y y + n_Z$ then the relations between 3-D motion + structure and disparity fields and relative flow fields can be established simply by:

$$\begin{cases} \frac{\Delta u(x_l,y_l,\delta)}{\delta(x_l,y_l)} &= \frac{T_Z}{n_Z} + (n_X T_Z - \omega_Y)x_l + (\omega_X + n_Y T_Z)y_l \\ \Delta v(x_l, y_l, \delta) &= 0 \end{cases} \tag{62}$$

In order to avoid bias in the estimation of initial 2-D motion fields, the latter are filtered by adapted filters (radial flow filtering for the relative flow, 2nd order filtering for the fields themselves) [53]. The 3-D motion estimation method proceeds in accordance with the following principles:

- stage 1: estimation, segmentation and filtering of 2-D apparent motion fields
- stage 2: matching of primitives based on coherence equations (62)
- stage 3: use of disparity functions for the reconstitution of surfaces between discontinuity regions detected during monocular analysis (stage 1)
- stage 4: estimation of 3-D motion parameters

A temporal linking phase is also introduced in order to allow a "sub-pixel" accuracy in the estimated disparity field (by temporal interpolation) and tracking along the temporal axis of discontinuity regions and matched segmented regions.

3.4 Joint motion estimation under stereoscopic constraints

In several applications - and notably those of stereoscopic sequence coding, where 3-D reconstruction is not an aim - it is sometimes not necessary to go back as far as the estimation of explicit 3-D motion and structure parameters. *A contrario*, it would appear interesting to move on to the 2-D or $2\frac{1}{2}$-D motion descriptor estimation phases not independently of each stereoscopic sequence, but jointly by introducing stereoscopic constraints into the estimation schemes themselves, linking the two descriptor fields.

In the case where only dense 2-D primitive fields are estimated (disparity fields $\{\delta_t\}$ and $\{\delta_{t+1}\}$ and apparent motion fields $\left\{\Theta_l = \vec{d_l}\right\}$, $\left\{\Theta_r = \vec{d_r}\right\}$) an available coherence constraint for these fields is to impose, at each point of the image plane, a linear relation:

$$\vec{d_l} + \vec{\delta_l} + \vec{d_r} + \vec{\delta_r} = 0 \tag{63}$$

consisting of forcing the closure of the quadrilateral illustrated in Figure 13.

Such a relation makes it possible, knowing three information fields, to deduce the fourth, an ability which is easily applied in the case where, given that the dense disparity fields are calculated on each stereoscopic pair, the knowledge of a motion field (for example on the left sequence) makes it possible to deduce the other field (on the right sequence). Tamtaoui and Labit [46] tested this estimation approach. It turns out that this too localized and too major constraint, notably on occlusion regions, can only provide an initial prediction of a field which then has to be affine to obtain results in motion compensation identical to the monocular case; obviously, this post-processing removes the previous stereoscopic constraint. Furthermore, this scheme remains very sensitive to the estimation bias of each of the information fields introduced.

An interesting alternative [46], [48] is to begin with a coherence equation linking the apparent motion fields $\vec{d_l} = (u_l, v_l)^t$ and $\vec{d_r} = (u_r, v_r)^t$ under stereoscopic constraints. This relation establishes itself as follows: if

$$\begin{bmatrix} X_r \\ Y_r \\ Z_r \end{bmatrix} = \mathrm{R}_l^r \begin{bmatrix} X_l \\ Y_l \\ Z_l \end{bmatrix} + \mathrm{T}_l^r \tag{64}$$

with $\mathrm{T}_l^r = (t_1, t_2, t_3)^t$ and $\mathrm{R}_l^r = (\ r_{i,j}\ )$ with $i = 1, 2, 3\ ,\ j = 1, 2, 3$, and if we assume that $Z_l = Z_r$ for all matched pixels (parallel cameras hypothesis), then it is possible to establish the following relation between apparent 2-D motion fields:

$$(r_{21} - \frac{t_2}{t_1} r_{11}) u_l + (r_{22} - \frac{t_2}{t_1} r_{12}) v_l = -\frac{t_2}{t_1} u_r + v_r \tag{65}$$

which can be put in the form $\alpha u_l + \beta v_l + \gamma u_r + \delta v_r = 0$ with:

$$\begin{cases} \alpha &= \frac{r_{21}}{t_2} - \frac{r_{11}}{t_1} \\ \beta &= \frac{r_{22}}{t_2} - \frac{r_{12}}{t_1} \\ \gamma &= \frac{1}{t_1} \\ \delta &= -\frac{1}{t_2} \end{cases} \tag{66}$$

It is equivalent matrically to $C\ .\ \Psi = 0$ with:

- $\Psi = (u_l, v_l, u_r, v_r)^t$ the motion vector linked to the two stereoscopic sequences,
- C coherence coefficients.

Tamtaoui and Labit [46] introduce this coherence equation within a pel-recursive type estimation scheme by minimization of a reconstitution error quadratic function (Γ) linked to the left and right sequences by gradient techniques. Namely:

$$\Gamma(\Psi, p_{lr}) = DFD^2(p_l, \vec{d_l}) + DFD^2(p_r, \vec{d_r}) \tag{67}$$

with p_{lr}, a couple of pixels (p_l, p_r) matched together; the estimation algorithm is then written:

$$\Psi^{k+1} = \Psi^k - \epsilon P \nabla_\Psi \Gamma(\Psi^k) \tag{68}$$

with $P = \mathbb{I} - C^T (C C^T)^{-1} C$. The matrix P is the matrix of projection on the coherence space:

$$\left\{ \Psi \in \mathbb{R}^4 \ / \ C.\Psi = 0 \right\} \tag{69}$$

This estimation technique (see Figure 16) compares favourably with monocular independent motion estimation techniques (see Figure 15) and with disparity estimation techniques (see Figure 14) used for compensation schemes.

Naturally, this approach on a dense field extends to region motion descriptor estimation methods (see Figure 17) by the use of parametric motion models [47]. In addition to the more global nature of these descriptors, such an approach appears more robust to estimation bias on the disparity since in this context it is a matter of matching regions and not points.

Some results below illustrate the performances achieved using these joint estimation algorithms concerning quality criteria of reconstitution after motion compensation and quality criteria of motion fields obtained.

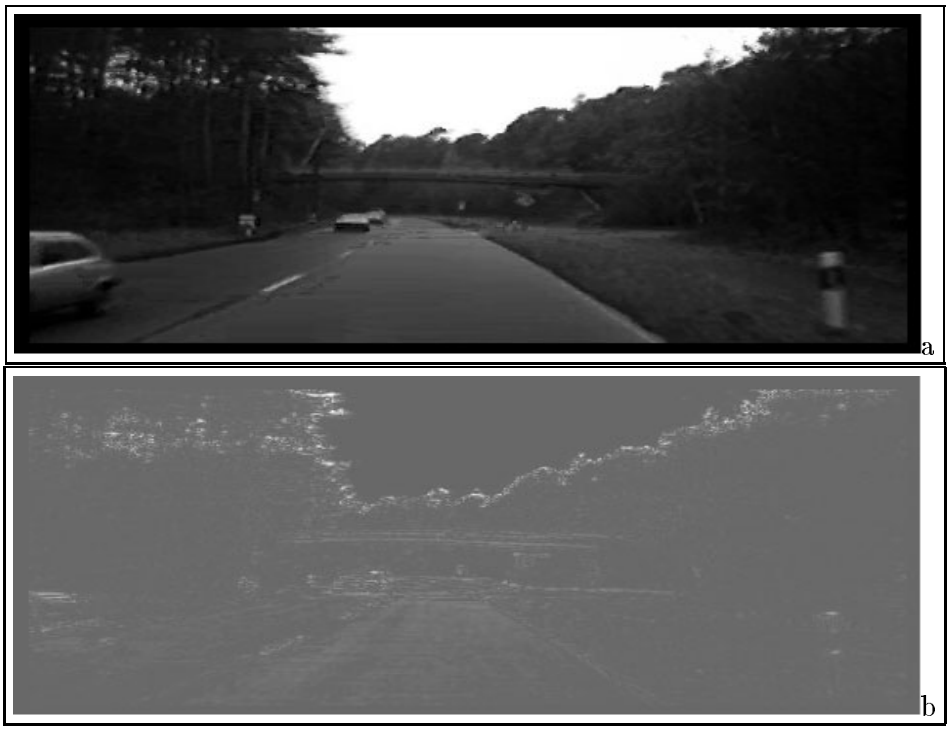

Figure 14: (a) Reconstructed "Campagne" image using disparity compensation, (b) Corresponding disparity compensation errors (MSE=54.24)

3.5 Application to coding of stereoscopic sequences (3-D TV)

3.5.1 The general context of 3-D TV

As Figure 18 illustrates, a three-dimensional television system (3-D TV) consists of various elements as follows:

- a stereoscopic capture system (at least two cameras, calibrated or not)
- a coder-decoder implementing a compression phase for transmission or storage of stereoscopic sequences
- a 3-D display for which various technologies exist: dual-screens with polarizing filters, glasses with synchronized obturators, lenticular plate screens,...

The motion estimation algorithms using stereovision-motion cooperation, mentioned in the previous paragraphs, integrate naturally into such an applicational context in order to analyze stereoscopic source-sequences and code them by motion and/or disparity compensation.

3.5.2 Stereoscopic sequence coding strategies

We remain within the context of compatible coding-decoding-restitution approaches, *i.e.*, which permit restoration of a monocular view, if the receiver does not have a 3-D display.

Two definitions of compatibility can then be introduced (see Figure 19):

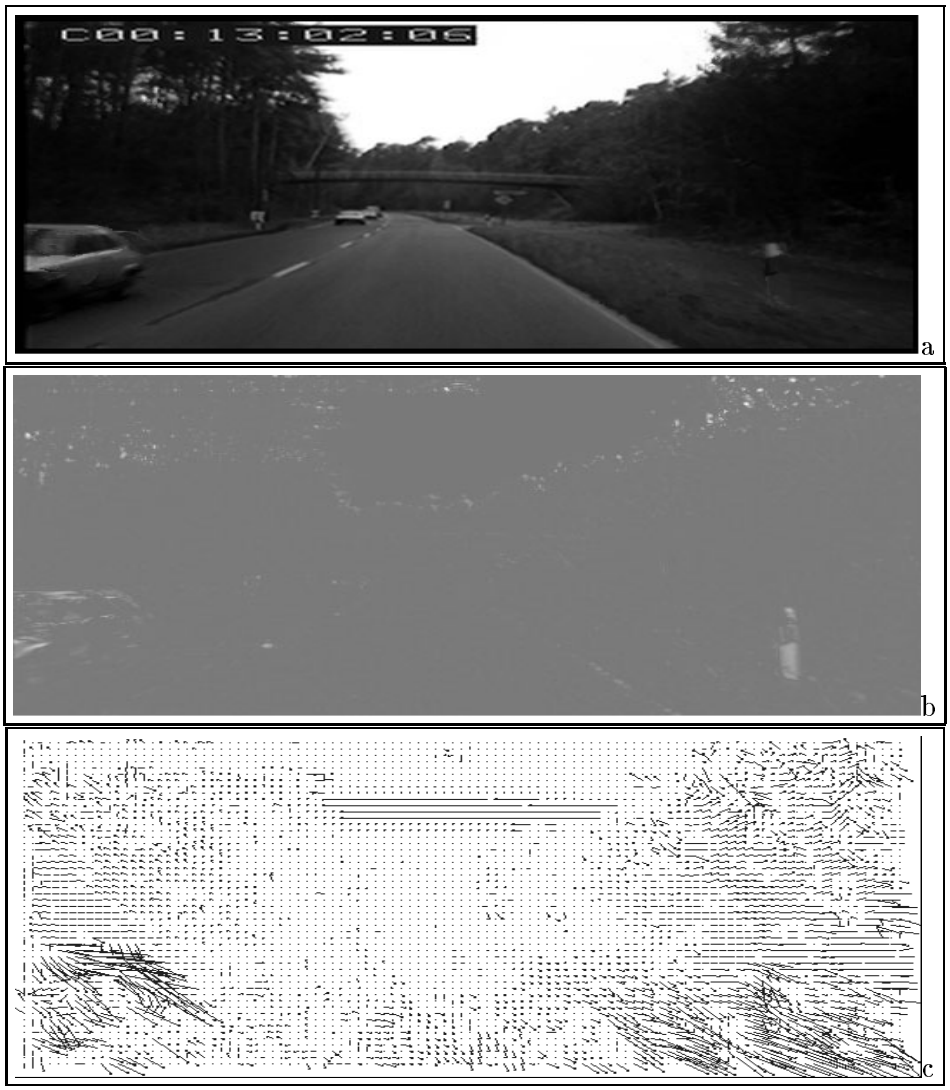

Figure 15: (a) Reconstructed "Campagne" image using motion compensation (Walker-Rao pel-recursive method), (b) Corresponding motion compensated errors (MSE=7.92), (c) Motion vector field

1. in the first approach, we assume the coding of one of the stereoscopic sequences (for example the left as illustrated in Fig 19) by such a standard monocular sequence compression technique. The second sequence will be coded by:

   - disparity compensation [57] (example in Figure 14)
   - motion compensation [47], [10] (examples in Figures 15 to 17).

   The second coding channel is thus used to transmit compensation errors and if necessary, if the disparity and motion information fields are used non-predictively in the compensation scheme, these should also be transmitted.

   In this case, an effective stereo-motion cooperation approach makes it possible:

   - to compare the two possible types of compensation

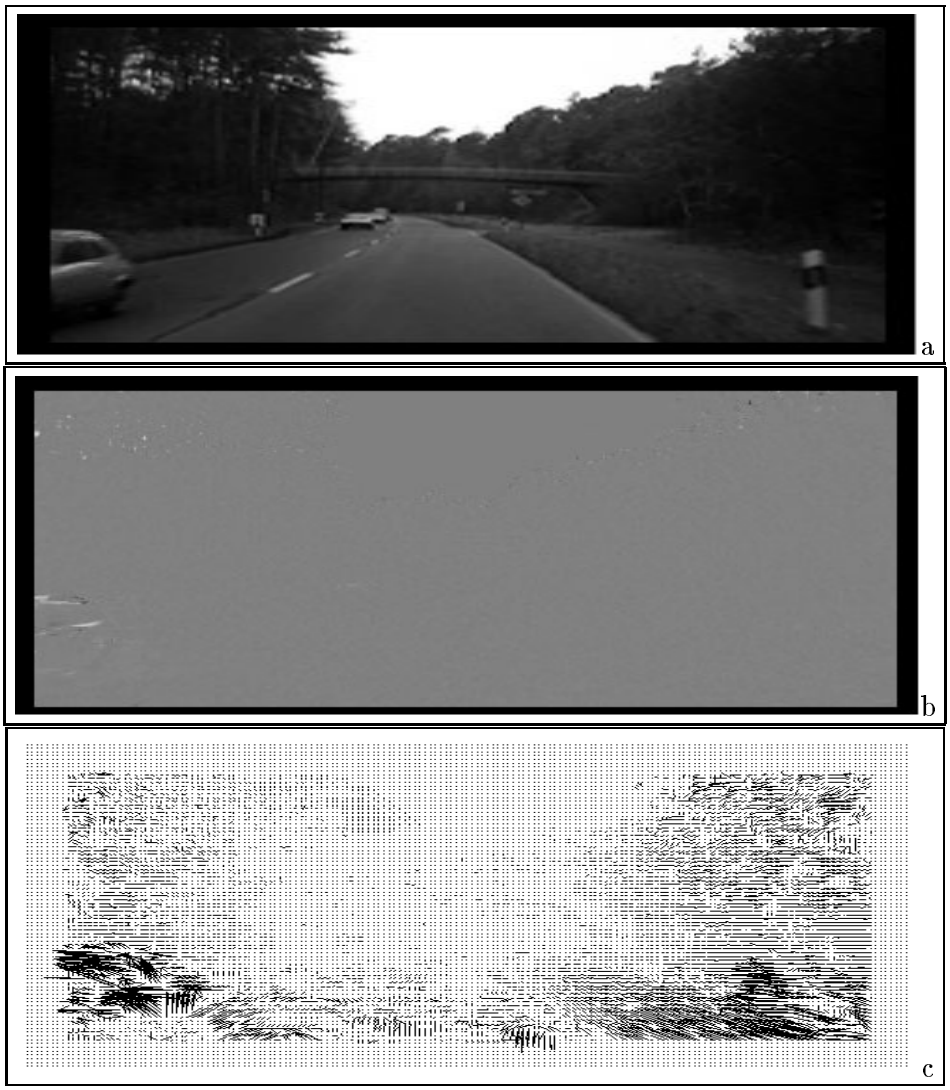

Figure 16: (a) Reconstructed "Campagne" right image using joint coherent motion compensation on the two stereoscopic sequences, (b) Corresponding motion compensated errors (MSE=3.73), (c) Motion vector field

- to restrict the volume of information which represents these fields by taking account of equations of geometric dependence which link them (coherence equations described just before)
- to minimize depth perception artefacts which are linked to an independent view-to-view reconstitution by purely monocular approaches.

2. the second approach appears as an attractive, but more difficult to achieve, extension of the previous notion of compatibility. Prior to any coding of stereoscopic sequences, a joint stereo-motion analysis is carried out. From this processing phase are generated, on the one hand, a "compatible" monocular sequence which can be situated as an intermediate position between the viewpoints of the left and right cameras and, on the other hand, innovation information (identical in nature to the compensation error information previously described) with regard to this compat-

Figure 17: (a) Reconstructed "Campagne" right image using joint coherent quadtree-based affine motion estimation on the two stereoscopic sequences, (b) Corresponding motion compensated errors (MSE=15.16), (c) Motion vector field

ible sequence. Such an approach is well adapted to the case of the use of 3-D motion+structure estimation methods which, once carried out, make it possible to synthesize the 3-D scene perceived from all viewing angles. This coding strategy, difficult because of the even more imprecise nature of 3-D parameter estimations obtained on true stereoscopic sequences, can be considered as a natural extension of the Analysis-Synthesis or object-oriented coding approaches, described in paragraph 8.2.6 for simple objects.

References

[1] G. Adiv, "Determining three-dimensional motion and structure from optical flow generated by several moving objects", *IEEE Trans. on Pattern Analysis and Ma-*

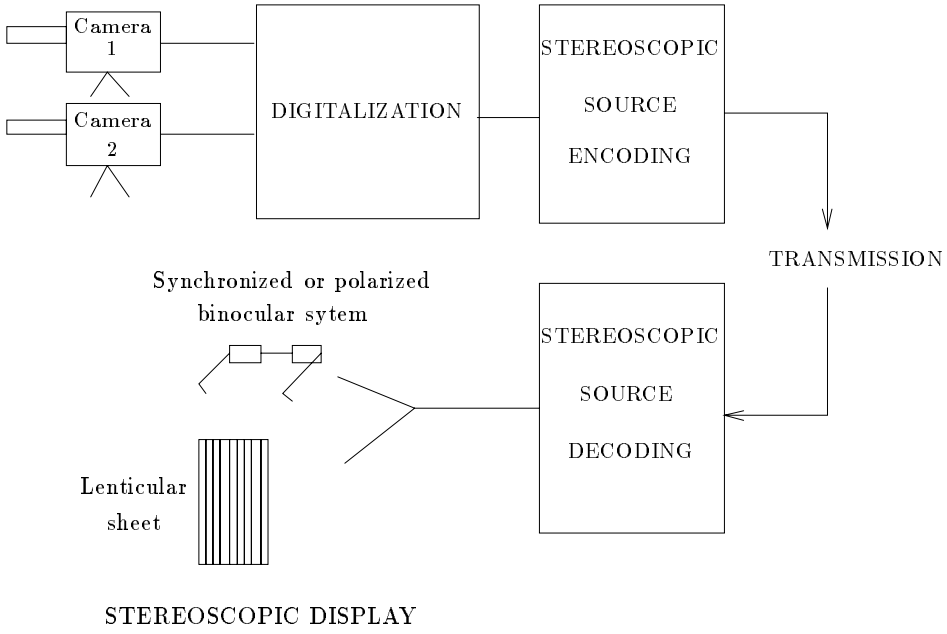

Figure 18: General scheme of a 3-D TV system

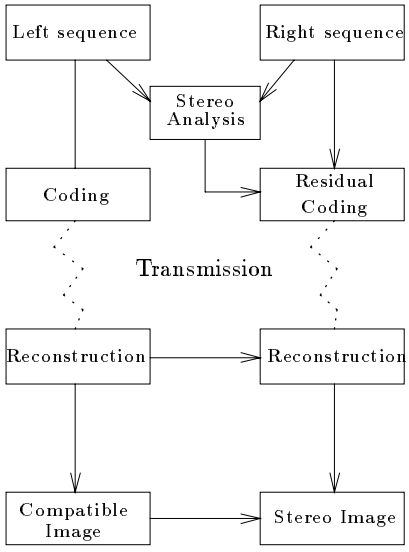

Figure 19: Compatibility approach for transmission of stereoscopic image sequences

chine Intelligence, Vol. PAMI-7, pp. 384-401, July 1985.

[2] G. Adiv, "Inherent ambiguities in recovering 3D motion and structures from a noisy flow field", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-11, pp. 477-489, May 1989.

[3] K. Aizawa, H. Harashima, and T. Saito, "Model-based analysis synthesis image coding (MBASIC) system for a person's face", *Signal Processing: Image Communication*, Vol. 1, pp. 139-152, 1989.

[4] P. Anandan, "A unified perspective on computational techniques for the measurement of visual motion", *Proc. of the 1st Int. Conf. on Computer Vision*, pp. 219-230, May 1987.

[5] J.L. Barron, A.D. Jepson, and J.K. Tsotsos, "The feasibility of motion and structure from noisy time-varying image velocity information", *Int. Journal of Computer Vision*, pp. 239-269, 1990.

[6] P. Bouthemy and J. Santillana-Rivero, "A hierarchical likelihood approach for region segmentation according to motion-based criteria", *Proc. of the 1st Int. Conf. on Computer Vision*, London, pp. 463-467, 1987.

[7] N. Diehl, "Object-oriented Motion Estimation and segmentation in Image Sequences", *Signal Processing: Image Communication*, Vol. 3, No. 1, pp. 23-56, 1991.

[8] E. Dubois, "Motion-compensated filtering of time-varying images", *Multidimensional Systems and Signal Processing*, No. 3, pp. 211-239, 1992.

[9] J.L. Dugelay and B. Choquet, "A 3D image analysis algorithm and stereoscopic television", *Proc. of Festival Int. des images 3D*, Paris, Sept. 1991.

[10] J.L. Dugelay and D. Pele, "Motion and disparity analysis of a stereoscopic sequence: application to 3DTV encoding", *European Conference on Signal Processing, EUSIPCO'92*, Aug. 1992.

[11] R. Forchheimer and O. Fahlander, "Low bit rate coding through animation", *Picture Coding Symposium, PCS'83*, Davis, March 1983.

[12] R. Forchheimer, O. Fahlander, and T. Kronander, "A semantic approach to the transmission of face images", *Picture Coding Symposium, PCS'84*, Rennes, July 1984.

[13] E. François and P. Bouthemy, "The derivation of qualitative information in motion analysis", *Proc. of the 1st European Conf. on Computer Vision, ECCV'90*, pp. 226-230, 1990.

[14] E. François, *Interprétation qualitative du mouvement à partir d'une séquence d'images*, Ph-D thesis, Université de Rennes-I, June 1991.

[15] R. Hartley, "Segmentation of optical flow fields by pyramid linking", *Pattern Recognition Letters*, Vol. 3, pp. 253-262, July 1985.

[16] M. Hoetter, "Differential estimation of the global motion parameters zoom and pan", *Signal Processing*, Vol. 16, pp. 249-265, 1989.

[17] B.K.P. Horn and B. Schunck, "Determining optical flow", *Artificial Intelligence*, Vol. 17, pp. 185-203, 1981.

[18] B.K.P. Horn and J.R. Weldon, "Direct methods for recovering motion", *Int. Journal of Computer Vision*, Vol. 2, pp. 51-76, 1988.

[19] M. Hötter, "Object-oriented analysis-synthesis coding based on moving two-dimensional objects", *Signal Processing: Image Communication*, Vol. 2, pp. 409-429, 1990.

[20] M. Kanado, A. Koike and Y. Hatori, "Codings with knowledge-based analysis of motion pictures", *Picture Coding Symposium, PCS'87*, Stockholm, June 1987.

[21] Y.C. Kim and J.K. Aggarwal, "Determining object motion in a sequence of stereo images", *IEEE Journal of Robotics and Automation*, Vol. 3, No. 6, pp. 599-614, Dec. 1987.

[22] C. Labit and H. Nicolas, "Compact motion representation based on global features for semantic image sequence coding", *Proc. of the SPIE Conf. on Visual Communication and Image Processing, VCIP'91*, Vol. 2, pp.697-709, Nov. 1991.

[23] M. K. Leung and T. S. Huang, "An integrated approach to 3D motion analysis and object recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-13, No. 10, pp. 1075-1084, Oct 1991.

[24] S. X. Li and M. H. Loew, "The quadcode and its arithmetic", *Communications of the ACM*, pp. 621-631, July 1987.

[25] L. Lingxiao, T. S. Huang *et al.*, "Motion estimation from 3-D points sets with and without correspondences", *Proc. of the Conf. Computer Vision and Pattern Recognition, CVPR'86*, pp. 194-201, 1986.

[26] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections", *Nature*, Vol. 293, pp. 133-135, Sept. 1981.

[27] A. Mitiche and P. Bouthemy, "Tracking modelled objects using binocular images", *Computer Vision, Graphics and Image Processing*, Vol. 32, pp. 384-396, 1985.

[28] A. Mitiche, "On kineopsis and computation of structure and motion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 1, pp. 109-112, Jan. 1986.

[29] Y. Miyamoto and M. Ohta, "Global motion compensation for rotation and zooming image", *Proc. of the Picture Coding Symposium, PCS'91*, pp. 137-140, Sept. 1991.

[30] H.-G. Musmann, M. Hötter and J. Ostermann, "Object-oriented analysis-synthesis coding of moving images", *Signal Processing: Image Communication*, Vol. 1, pp. 117-138, 1989.

[31] N. Navab, Z. Zhang and O. D. Faugeras, "Tracking, motion and stereo", *Proc. of the Scandinavian Conf. on Image Analysis, SCIA'91*, pp. 98-105, 1991.

[32] S. Negahdaripour and A. Yuille, *Direct passive navigation, I: analytical solutions for planes*, AI Memo 863, MIT Artificial Intelligence Lab, August 1985.

[33] S. Negahdaripour and A. Yuille, "Direct passive navigation, II: analytical solutions for quadratic patches", *Conf. Computer Vision and Pattern Recognition, CVPR'88*, pp. 404-410, 1988.

[34] A.N. Netravali, T.S. Huang *et al*, "Algebraic Methods in 3D Motion Estimation from two-view point correspondences", *Int. Journal of Imaging Systems and Technology*, Vol. 1, pp. 78-99, 1989.

[35] A. N. Netravali and J. D. Robbins, "Motion compensated television coding: Part I", *Bell Syst. Tech. Journal*, Vol. 58, No. 3, pp. 631-670, March 1979.

[36] A. N. Netravali and J. Salz, "Algorithms for estimation of three-dimensional motion", *AT&T Technical Journal*, Vol. 64, No. 2, Feb. 1985.

[37] H. Nicolas and C. Labit, "Global motion identification for image sequence analysis and coding", *Proc. of Int. Cong. on Speech, Acoustics and Signal Processing, ICASSP'91*, Vol. 4, pp. 2825-2828, May 1991.

[38] H. Nicolas and C. Labit, "Region-based motion estimation using deterministic relaxation schemes for image sequence coding", *Proc. Int. Cong. on Speech, Acoustics and Signal Processing, ICASSP'92*, Vol. 3, pp. 265-268, March 1992.

[39] H. Nicolas, *Hiérarchie de modèles de mouvement et méthodes d'estimation associées. Application au codage de séquences d'images*, Ph-D Thesis, Université de Rennes-I, Sept. 1992.

[40] J. Rissanen, "Modeling by shortest data description", *Automatica*, Vol. 14, pp. 465-472, 1986.

[41] H. Samet, "Quadtree from boundary codes", *Communications of the ACM*, pp. 163-170, March 1980.

[42] H. Sanson, "Motion affine models identification and application to television image coding", *SPIE Conf. Visual Communication and Image Processing, VCIP'91*, Vol. 1605, pp. 570-581, Nov. 1991.

[43] J. Santillana-Rivero, P. Bouthemy and C. Labit, "Hierarchical motion-based image segmentation applied to HDTV", *2nd Int. Workshop on Signal Processing of HDTV*, l'Aquila, March 1988.

[44] P. Y. Simard and G. E. Mailloux, "A projection operator for the restoration of divergence-free vector fields", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-10, No. 2, pp. 248-256, 1988.

[45] M. Subbarao and A. M. Waxman, "On the uniqueness of image flow solutions for planar surfaces in motion", *Computer Vision, Graphics and Image Processing*, Vol. 36, pp. 208-220, 1986.

[46] A. Tamtaoui and C. Labit, "Constrained disparity and motion estimators for 3DTV image sequence coding", *Signal Processing: Image Communication*, Vol. 4, pp. 45-54, 1991.

[47] A. Tamtaoui, *Coopération stéréovision-mouvement pour la compression de séquences stéréoscopiques. Application à la Télévision en relief (TV3D)*, Ph-D Thesis, Université de Rennes-I, Oct. 1992.

[48] A. Tamtaoui and C. Labit, "Constrained motion estimators for 3D sequence coding", *Proc. of the European Conf. on Signal Processing, EUSIPCO'92*, Brussels, Aug. 1992.

[49] M. Tistarelli, E. Grosso and G. Sandini, "Dynamic stereo in visual navigation", *Proc. of Conf. Computer Vision and Pattern Recognition, CVPR'91*, pp. 186-192, 1991.

[50] Y. T. Tse and R. Baker, "Global zoom/pan estimation and compensation for video compression", *Proc. of Int. Cong. on Speech, Acoustics and Signal Processing, ICASSP'91*, Vol. 4, pp. 2725-2728, May 1991.

[51] A. Verri, F. Girosi and V. Torre, "Mathematical properties of the two-dimensional motion field: from singular points to motion parameters", *Journal of Optical Soc. of Am.*, Vol. 6, No. 5, pp. 698-712, May 1989.

[52] A. M. Waxman and K. Wohn, "Contour evolution, neighborhood deformation, and global image flow: planar surfaces in motion", *Int. Journal of Robotics Research*, Vol. 4, No. 3, pp. 95-108, 1985.

[53] A. M. Waxman and S. Sinha, "Dynamic stereo; passive ranging to moving objects from relative image flows", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 4, pp. 406-412, July 1986.

[54] A. M. Waxman and J. H. Duncan, "Binocular image flows: steps forward stereo-motion fusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 6, pp. 715-729, Nov. 1986.

[55] A. M. Waxman and K. Wohn, "Image flow theory: a framework for 3D inference from time-varying imagery", *Chapter 3 in Advances in Computer Vision*, Erlbaum Associates Ed., London, pp. 164-224, 1988.

[56] S. F. Wu and J. Kittler, "A differential method for simultaneous estimation of rotation, change of scale and translation", *Signal Processing: Image Communication*, Vol. 2, pp. 69-80, 1990.

[57] M. Ziegler, "Disparity estimation using variable blocksize", *Proc. of the 3rd COST-230 Workshop on 3DTV Signal Processing*, Rennes, 1992.