INTERPOLATIVE CODING OF IMAGE SEQUENCES USING TEMPORAL LINKING OF MOTION-BASED SEGMENTATION

Laurent Bonnaud [1] *Claude Labit* [1] *Janusz Konrad* [2]

[1] IRISA/INRIA-Rennes, Campus de Beaulieu, F-35042 Rennes Cédex, France, E-mail: <name>@irisa.fr
[2] INRS-Telecom, 16 Place du commerce, Verdun, Québec, Canada, H3E 1H6, E-mail: konrad@inrs-telecom.uquebec.ca

ABSTRACT

This paper presents a new temporal interpolation algorithm based on segmentation of images into polygonal regions undergoing affine motion. The goal of this work is to improve upon the block-based interpolation used in MPEG (B-frames). In the first part, we briefly describe the region-based framework and the temporal linking algorithm that jointly provide the segmentation and motion parameters. In the second part, we present various applications of the proposed algorithm to temporal interpolative prediction. We examine one of these schemes in detail, including the special processing of occlusion areas. Results are illustrated by predicted images and using the MSE criterion we compare their quality with other schemes.

1. INTRODUCTION

In order to exploit high temporal redundancy in image sequences, usually image prediction based on motion compensation is used. Motion-compensated prediction can be achieved using only a previous image (P-frames in MPEG) or both previous and following images (B-frames). In MPEG standard, a block-oriented translational motion model is used: the same motion vector is applied to all pixels in a 16 × 16 block. This fixed partition cannot handle areas of complex motion or occlusion boundaries, and creates visually disturbing blocking artefacts in the reconstructed image at low bit rates. Furthermore, the coding of motion vectors is inefficient since motion of large regions could be described with only a few motion parameters (in case of global pan or zoom, for instance). To achieve lower bit rates, region-oriented motion compensation has been introduced [1, 2]. Those schemes, however, use the previous image only, disallowing prediction of uncovered areas. Our new interpolation scheme is able to predict uncovered areas using the following image and is also capable of ensuring better prediction in other areas.

2. TEMPORAL LINKING OF THE SEGMENTATION

In this study, images are segmented into homogeneous regions. Regions have an arbitrary polygonal shape, and form a partition of the image. The homogeneity criterion is based on a 4-parameter simplified affine model or a 6-parameter complete affine model. The notation $\Theta^{\pm}_{\mathcal{R},t_i\to t_j}$ describes a motion descriptor from image I_{t_i} to image I_{t_j} for the region $\mathcal{R}$ with a $^+$ exponent if $t_i < t_j$ (forward direction) or a $^-$ exponent if $t_j < t_i$ (backward direction). For each point $p \in \mathcal{R}_{t_i}$ the displacement vector from image I_{t_i} to image I_{t_j} is defined as $\vec{d}^{\pm}_{t_i\to t_j}(p)$ (see figure 2).

In the simplified affine case, $\Theta^{\pm}_{\mathcal{R},t_i\to t_j} = [t_x, t_y, k, \theta]$ and

$$\vec{d}^{\pm}_{t_i\to t_j}(p) = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{bmatrix} k & -\theta \\ \theta & k \end{bmatrix} \begin{pmatrix} x_p - x_r \\ y_p - y_r \end{pmatrix}$$

where r is a reference point. It is defined as the center of gravity of the region $\mathcal{R}_{t_i}$ if it not being occluded.

In the affine case, $\Theta^{\pm}_{\mathcal{R},t_i\to t_j} = [t_x, t_y, a, b, c, d]$ and

$$\vec{d}^{\pm}_{t_i\to t_j}(p) = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x_p - x_r \\ y_p - y_r \end{pmatrix}.$$

When descriptor Θ^+ needs to be calculated from descriptor Θ^- (and conversely), inverse geometric transformation is computed by inversion of the 2 × 2 matrix.

As the estimation of motion parameters requires a good initialization and thus can be biased, a long-term filtering of motion parameters (based on temporal recursive Kalman filtering) has been implemented.

The temporal linking algorithm consists of 3 steps: for image I_t

- Motion parameters $\Theta^{+}_{\mathcal{R},t-\delta t\to t}$ and shape of regions are predicted with a Kalman filter (δt is the period of image acquisition). Each motion parameter is predicted independently according to a constant acceleration model as in [3, 4].
- Predicted spatio-temporal segmentation is adjusted with respect to actual edges in the image I_t using a snake approach: the minimized energy is the sum of $-\|\vec{\nabla} I_t\|$ along the polygonal region boundaries. Motion of the snake is constrained to be affine and the minimization is achieved using a gradient descent on motion parameters [5].
- Motion parameters $\Theta^{-}_{\mathcal{R},t\to t-\delta t}$ of each region are estimated using a region matching: the MSE is minimized by a gradient descent. As motion vectors are non-integer, MSE is computed with spatial bicubic interpolation [6] at inter-pixel positions. The gradient based minimization algorithm needs image derivatives ; to be consistent, they are computed with the same bicubic interpolation [7]. Four different motion descriptors are selected as an initialization and parameters giving the lowest MSE are chosen *a posteriori*. This choice is necessary to avoid a possible divergence of the Kalman filter. The four initial parameter sets are the null descriptor, parameters estimated for the same region in the previous image, filtered parameters and predicted parameters.

0-7803-2431-5/95 $4.00 © 1995 IEEE

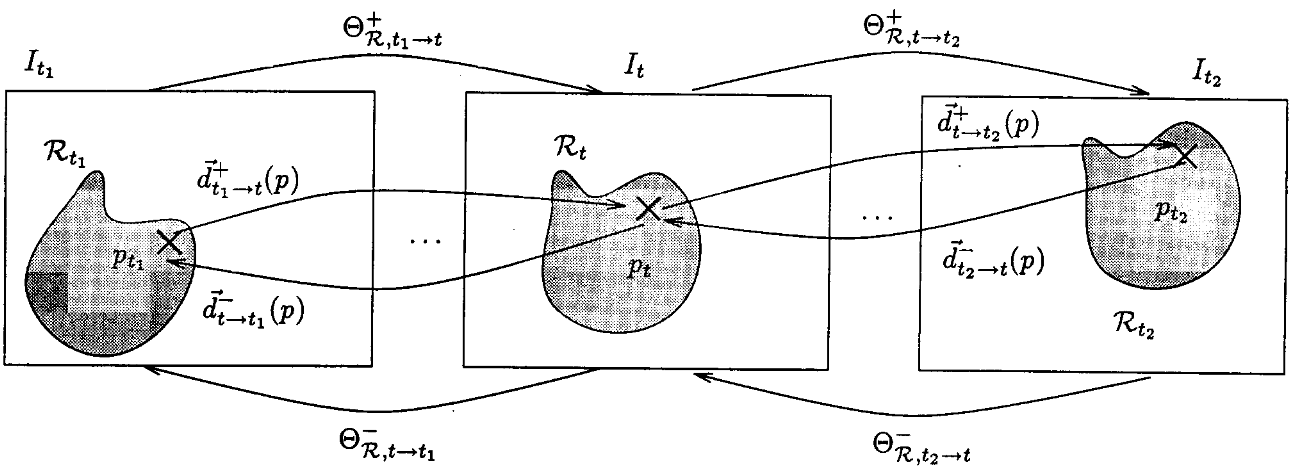
$\Theta^+_{\mathcal{R},t_1\to t}$
$\Theta^+_{\mathcal{R},t\to t_2}$
I_{t_1}
I_t
I_{t_2}
$\mathcal{R}_{t_1}$
$\vec{d}^+_{t_1\to t}(p)$
p_{t_1}
$\vec{d}^-_{t\to t_1}(p)$
$\mathcal{R}_t$
p_t
$\vec{d}^+_{t\to t_2}(p)$
p_{t_2}
$\vec{d}^-_{t_2\to t}(p)$
$\mathcal{R}_{t_2}$
$\Theta^-_{\mathcal{R},t\to t_1}$
$\Theta^-_{\mathcal{R},t_2\to t}$

Figure 2. Forward and backward motion vectors and descriptors.

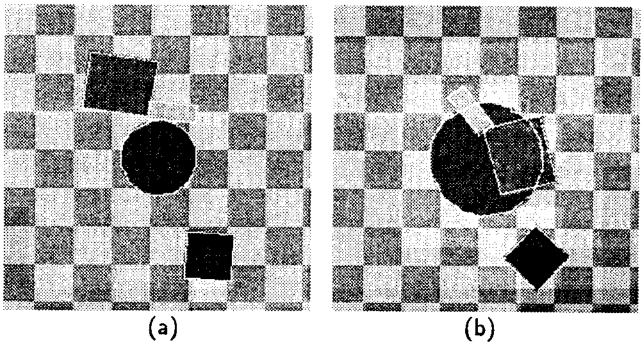

Figure 1. (a) Initialization of segmentation for the first frame of the sequence. (b) Segmentation of the last frame after processing the whole sequence (19 images).

The result of the segmentation process is shown in Figure 1. More details on the algorithm and on the experimental results can be found in [8].

3. REGION-BASED BIDIRECTIONAL MOTION COMPENSATION

The segmentation algorithm processes all images between two consecutive I- or P-frames called reference images (I_{t_1} and I_{t_2}). It provides motion descriptors of the form $\Theta^-_{\mathcal{R},t+\delta t\to t}$ between successive images, too. Then, the segmentation map and motion parameters are used to interpolate the intermediate B-frames. In order to predict the B-frames $(I_t)_{t_1<t<t_2}$, the decoder needs the segmentations of the three images and for each region $\mathcal{R}$ the two motion descriptors from the B-frame to the reference images, namely $\Theta^-_{\mathcal{R},t\to t_1}$ and $\Theta^+_{\mathcal{R},t\to t_2}$. The coder transmits the segmentation of I_{t_1} and I_{t_2} to the decoder, as well as $\Theta^-_{\mathcal{R},t_2\to t_1}$ (used if I_{t_2} is a P-frame). This descriptor is computed using motion estimator initialized with parameters of the affine transformation which is a combination of affine transformations $(\Theta^-_{\mathcal{R},t+\delta t\to t})_{t_1\leq t<t_2}$.

There is a trade-off between the quantity of motion and segmentation information transmitted to the decoder and the quality of the predicted B-frame. Here are several possibilities:

- **bidirectional motion compensation.** The coder transmits the segmentation of I_t, as well as $\Theta^-_{\mathcal{R},t\to t_1}$ and $\Theta^+_{\mathcal{R},t\to t_2}$.
- **bidirectional segmentation prediction.** The coder transmits only $\Theta^-_{\mathcal{R},t\to t_1}$ and $\Theta^+_{\mathcal{R},t\to t_2}$. The decoder reconstructs the segmentation of I_t by applying $\Theta^+_{\mathcal{R},t_1\to t}$ to the boundaries of $\mathcal{R}_{t_1}$ and $\Theta^-_{\mathcal{R},t_2\to t}$ to the boundaries of $\mathcal{R}_{t_2}$.
- **pure interpolation.** The coder transmits no segmentation or motion information. The decoder interpolates $\Theta^-_{\mathcal{R},t\to t_1}$ and $\Theta^+_{\mathcal{R},t\to t_2}$ from $\Theta^-_{\mathcal{R},t_2\to t_1}$ and previously transmitted motion descriptors given a suitable motion model (constant velocity or acceleration for example). It reconstructs the segmentation of I_t as above, too.

Only the first scheme has been implemented and tested, assuming a lossless transmission of segmentation and motion parameters.

The interpolative prediction of the B-frame $\hat{I}_t$ itself is done for pixel $p_t \in \mathcal{R}_t$ as follows

$$p_{t_1} = p_t + \vec{d}^-_{t\to t_1}(p_t), p_{t_2} = p_t + \vec{d}^+_{t\to t_2}(p_t)$$

$$\hat{I}_t(p_t) = \alpha_p I_{t_1}(p_{t_1}) + \beta_p I_{t_2}(p_{t_2})$$

with spatial bicubic interpolation [6, 7] used at inter-pixel positions in I_{t_1} and I_{t_2}.

In order to handle occlusions, a distinction is made between "normal" areas and areas covered or uncovered by the movement of another region. The decision is taken on a pixel per pixel basis according to the rules in Table 1. Figure 3 illustrates the different types of areas for non-overlapping motion, but everything still holds for overlapping motion.

type of area and (α_p,β_p)	$p_{t_2} \in \mathcal{R}_{t_2}$	$p_{t_2} \notin \mathcal{R}_{t_2}$
$p_{t_1} \in \mathcal{R}_{t_1}$	"normal" area (α,β)	covered area $(1,0)$
$p_{t_1} \notin \mathcal{R}_{t_1}$	uncovered area $(0,1)$	spatial prediction $(0,0)$

Table 1. Classification of different types of areas and associated interpolation coefficients.

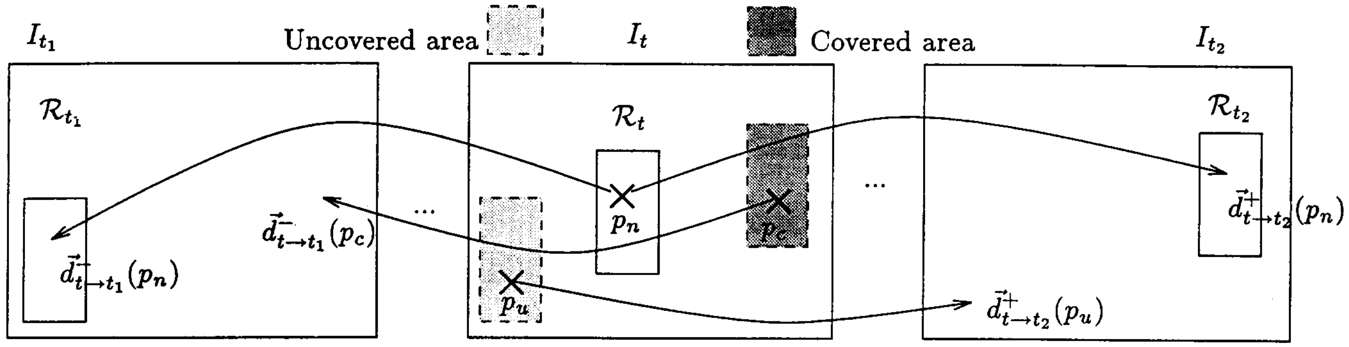

Figure 3. Interpolation of "normal" areas and of areas covered and uncovered by the movement of the region $\mathcal{R}$.

In the "normal" case, the interpolation is simply a linear combination of the corresponding points in the reference images with fixed coefficients α and β.

- If both motions are of the same quality and if there is no illumination variation, the simplest coefficients $(\alpha, \beta) = (0.5, 0.5)$ can be chosen.
- If there is an illumination variation, a model which is linear in time can be used: $(\alpha, \beta) = (\frac{t_2 - t}{t_2 - t_1}, \frac{t - t_1}{t_2 - t_1})$.
- Coefficients (α, β) can also be estimated at the same time as motion descriptors as in [9]. In this case, they have to be transmitted, too.

Covered and uncovered pixels are predicted by their intensity in the reference image where they appear.

The "singular" case $(\alpha_p, \beta_p) = (0, 0)$ happens only when p has no corresponding pixel in the reference images. Thus, temporal interpolation is not possible. Instead, spatial interpolation based on median filtering with a growing window is used. In order to ensure temporal coherence between successive spatial interpolations, median filtering takes into account only those pixels that belong to the same region $\mathcal{R}_t$ as p.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

We tested our region-based bidirectional prediction on a synthetic image sequence with three interpolated images between two I- or P-frames. We tried both fixed coefficients $(\alpha_p, \beta_p) = (0.5, 0.5)$ and time-linear coefficients, and made a comparison with both region-based monodirectional prediction and block-matching bidirectional prediction.

A comparison using the mean square error defined as

$$MSE(\mathcal{R}_t) = \frac{1}{\#\mathcal{R}_t} \sum_{p \in \mathcal{R}_t} [I_t(p) - \hat{I}_t(p)]^2$$

is shown in Figure 4. Bidirectionally interpolated images (FC-B-RB-MC) have a significantly lower MSE in comparison with images computed using previous-image prediction (M-RB-MC); 20–30 instead of 40–50. The energy (or innovation in uncovered areas) is concentrated in the next P-frame whose MSE is therefore higher. With such a low MSE (and good visual quality, see next paragraph), a hybrid coder could transmit prediction errors only for P-frames, thus achieving a coding gain. For B-frames our region-based prediction does as well as block-matching whereas for P-frames it is worse. This can be explained by the fact that our algorithm detects uncovered areas and does nothing to predict them (they stay black). The same kind of spatial interpolation as for our bidirectional prediction could be used, but the problem is more difficult because uncovered areas are larger ($4\delta t$ separate a P-frame and the image it is predicted from). For our particular sequence where motion estimates are correct, fixed coefficients give about the same result as coefficients that vary linearly over time.

Examples of interpolated images are shown in Figure 5. As can be seen, our algorithm has a drawback in the way it handles edges. Images often have somewhat blurred edges, because of prefiltering before sampling in real images and because of interpolation in synthetic images. Our segmentation does not take this into account and region boundaries cut the image abruptly. This explains why we can see "ghost edges" in interpolated images (for instance the background has a remaining edge of the dark rectangle). To correct this, we erode the segmentation mask of regions in the reference frames before we test if p_{t_1} or p_{t_2} belong to them. Therefore, the transition zone between the grey levels of regions across a boundary is not used in the interpolation. This causes more "singular" cases where $(\alpha_p, \beta_p) = (0, 0)$, but as long as the erosion is not too strong, the spatial prediction can handle those pixels. The result does not show the same blocking artefacts as block-matching.

We have described a new region-based interpolation scheme which is able to predict uncovered areas. Furthermore, predicted images have a significantly lower MSE than with a mono-directional prediction and are visually better than with block-matching. Work is currently in progress to test this interpolation scheme on real image sequences.

REFERENCES

[1] V. Garcia-Garduño, C. Labit, and L. Bonnaud. – Temporal linking of motion-based segmentation for object-oriented image sequence coding. – In *Proceedings of EUSIPCO 94*, University of Edinburgh, Scotland, UK, September 1994.

[2] H. Musmann, M. Hötter, and J. Ostermann. – Object-oriented analysis-synthesis coding of moving images. – *Signal Processing, Image Com.*, 1(2):117–138, 1989.

[3] F. Meyer. – Region-based tracking in an image sequence. – Technical Report 1723, INRIA, France, July 1992.

[4] F. Meyer. – Region-based tracking in an image sequence. – In G. Sandini, editor, *Proc. of Second European Conference on Computer Vision ECCV-92*, pages 476–484, Santa Margherita, Italy, July 1992. Springer-Verlag.

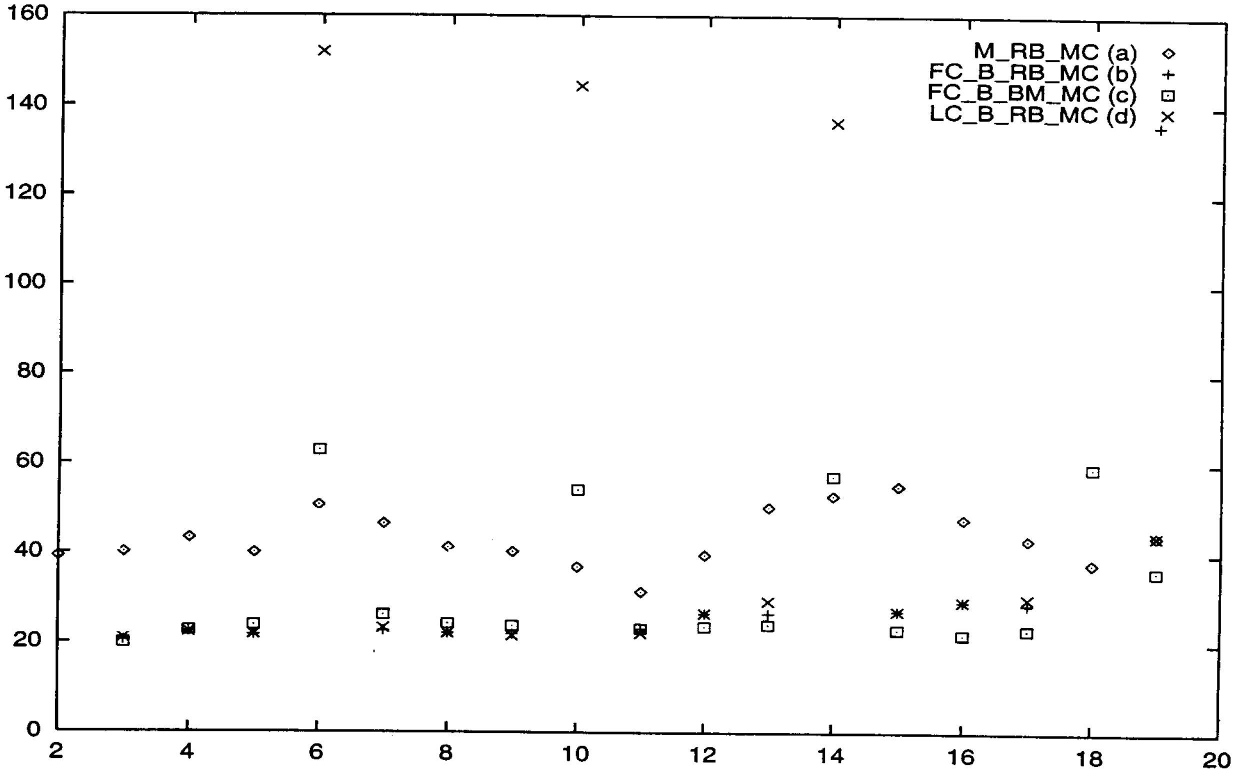

Figure 4. MSE for motion-compensated interpolation obtained with: (a) Monodirectional Region-Based Motion Compensation. (b) Fixed-Coefficients Bidirectional Region-Based MC. (c) Fixed-Coefficients Bidirectional Block-Matching MC. (d) Linear-Coefficients Bidirectional Region-Based MC. Images number 2, 6, 10, 14, 18 and 19 are P-frames ; other images are B-frames.

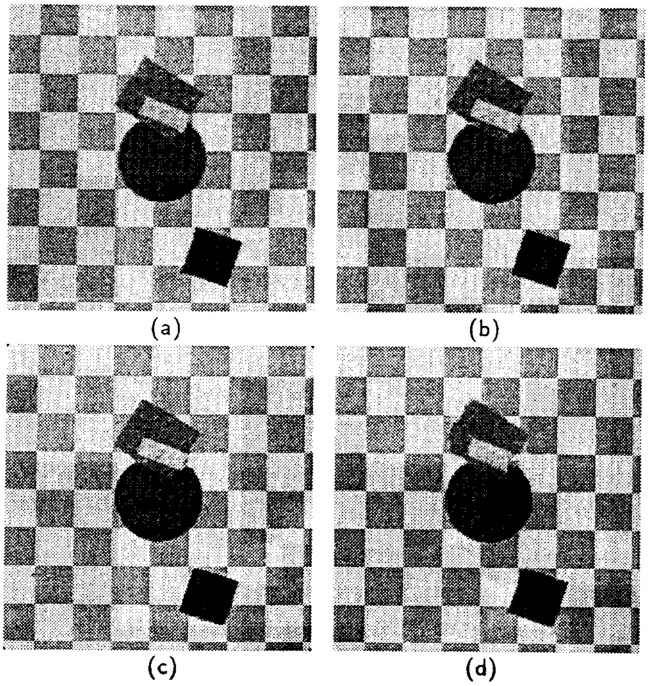

Figure 5. (a) Image number 8. (b) Region based interpolation. (c) Interpolated image with corrected edges. (d) Interpolated image with block-matching.

[5] B. Bascle and R. Deriche. – Features extraction using parametric snakes. – In *Proceedings of 11th IAPR Int. Conf. on Pattern Recognition (ICPR'92)*, volume 3, pages 659–662, The Hague, The Netherlands, August 1992.

[6] R.G. Keys. – Cubic convolution interpolation for digital image processing. – *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-29(6):1153–1160, December 1981.

[7] J. Konrad. – *Bayesian estimation of motion fields from image sequences.* – PhD thesis, McGill University, Dept. Electr. Eng., June 1989.

[8] L. Bonnaud. – Étude d'algorithmes de suivi temporel de segmentation basée mouvement pour la compression de séquences d'images. – Technical Report 2253, INRIA, France, January 1994. – ftp.inria.fr: INRIA/tech-reports/RR/RR-2253.ps.gz.

[9] H. Nicolas, J. Konrad, and C. Labit. – Joint estimation of motion and illumination variations for coding of image sequences. – In *Proc. Scandinavian Conf. Image Analysis*, May 1993.

[10] C. Bergeron and E. Dubois. – Parametric block estimation of motion and application to temporal interpolation of video sequences. – In *Proc. IEEE Int. Conf. Pattern Recognition*, pages 140–146, June 1990.

[11] G. Tziritas and C. Labit. – *Motion analysis for image sequence coding*, chapter 7, pages 269–285. – Elsevier, 1994.