SCALABLE REGION-MATCHING MOTION ESTIMATION BASED ON AN UNSUPERVISED SPATIAL SEGMENTATION

Brault P.[1], Mohammad-Djafari A.[2]
(1) Fundamental Electronics Institute, France
(2) Laboratory of Signals and Systems, France
(e-mail: patrice.brault@ief.u-psud.fr, http://braultp.free.fr)

Abstract

In a video scene, motion estimation (ME) can be studied on a dense field (optical flow) or on image structures or regions. Image structures can be deduced from the motion itself or formerly deduced by a segmentation. A scheme of ME, funded on a Bayesian segmentation using a Potts-Markov model, has lead to a "region-matching" ME scheme [5]. Bayesian segmentation has been operated, with the same model, in the wavelet domain and has shown an interesting gain in segmentation speed [6]. In the present work we have synthesized both approaches to demonstrate a new scheme of region-matching ME which uses the hierarchical property of the multiscale segmentation scheme. A bottom-top ME is built from the hierarchical segmentation in the wavelet domain. We show that a hierarchical, region-based, ME, can provide an interesting approach w.r.t. the necessity of ME robustness as well as its scalability, in a region (or object) -based compression scheme. This approach is to be compared with recent developments like the "structure from motion" (SfM) in [15], based on Bayesian inference and sequential Monte Carlo methods, and the "trace model" for object (face) detection and tracking in [12] (see also [13]) .

Key Words: Bayesian segmentation, Potts-Markov modeling, orthogonal wavelets multiresolution, motion estimation (ME), hierarchical ME, motion vectors (MVs), scalable ME, video compression, region matching..

1 Introduction

We investiguate here a Scalable Hierarchical Motion Estimation (SHME) for video compression. This SHME relies on a multiresolution, region-based, segmentation. The goal of this work is to show that, in a region-based approach of the segmentation, and of the ME, this one can be transmitted progressively. This method thus affords a real "scalability" and, hence, the ability to fit to low complexity decoders, which is not the case for yet studied hierarchical approaches [3, 1]. In the same way, if enables to compensate faster, and to decode, the spatially coded frames. The multiresolution construction of the ME provides a more robust estimation and decoding. The transmission of MVs at each scale and in an object approach is, to our knowledge, a very new scheme. It goes on the way of today's developments about semantic-based video coding, analysis and data mining.

The emergence of content-based representations of a video scene, like the "MPEG-4 visual" standard [16], has put aside pixel-based and block-based traditional scene representations. A content-based representation of a scene is mostly based on a partition of each frame, with an

interdependency of a number of successive frames [17, 14]. Algorithms following this strategy of scene representation can be classed as second generation coding algorithms. These algorithms use either of two different homogeneity criterions : the spatial or the temporal one. Some use both [17] .
Salembier [17] has studied a segmentation-based video coding with object manipulation. In his model, the ME of each region of the partition tree has to be estimated for the merging steps and for the decision. The motion of each region is represented by a polynomial model. Parametric models have been investigated [18, 9], but a simple affine model is used by Salembier. We have also investigated such a "trajectory approach" for ME in [7], but we do not develop it in this paper.

2 Incremental segmentation in the direct domain

Our ME scheme was initially based on the incremental unsupervised segmentation of a still image. This segmentation is operated in a Bayesian framework, and is based on a Potts-Markov model (1) in the pixel domain of a still image. This model was developed initially by Féron et al. and has been improved recently in [10]. The hypotheses made for the segmentation are : a gaussian noise ($\mathcal{N}_1$) over the whole image, a gaussian law ($\mathcal{N}_2$) for the pixels intensity in each region, an independency of the pixels in different regions, and the "attractive-repulsive" Potts model for the construction of the regions. In this section we recall the simple and efficient improvement brought in [5] by making the hypothesis that a close correlation can be assumed between the successive frames (images) of a sequence. This hypothesis leads to segment only the first frame of a sequence with a high number of iterations. The segmentation of the next frames is then initialized by the segmentation of the previous frame. This enables to drastically reduce the number of iterations and to increase the segmentation speed of the whole sequence.

$$p(z(\boldsymbol{r}), \boldsymbol{r} \in \mathcal{R}) = \frac{1}{T(\alpha)} \exp\left\{ \alpha \sum_{\boldsymbol{r} \in \mathcal{R}} \sum_{\boldsymbol{s} \in V(\boldsymbol{r})} \delta(z(\boldsymbol{r}) - z(\boldsymbol{s})) \right\} \tag{1}$$

PMRF with a first order neighborhood in the pixels domain

where $\boldsymbol{r}$ stands for the pixel position, z for the hidden Markov variable for the segments, s for the sites of a first order neighborhood, δ is the Kronecker symbol whose value stands for the potential energy, α is the Potts "attractivity" parameter and $T(\alpha)$ is a normalization coefficient.
From the knowledge of the observable $\underline{\mathbf{g}}$ and from the hypotheses formulated above, we are able to give an explicite expression of the posterior probability for the initial $\underline{\mathbf{f}}$ image, as well as for its segmentation $\underline{\mathbf{z}}$. For the computation of the MAP estimate, we use a Monte Carlo framework [8, 2]. This gives us, through the Gibbs sampling algorithm (2), the required numerical values to compute the estimate of the original image $\underline{\mathbf{f}}$ and of the segmented image $\underline{\mathbf{z}}$, as well as the hyperparameters (means and variances of the prior laws) knowing all the pixels of the observed image.

$$\left\{ \begin{array}{lcl} \boldsymbol{f}^n & \sim & p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{z}^{(n-1)}, \boldsymbol{\theta}^{(n-1)}) \\ \boldsymbol{z}^n & \sim & p(\boldsymbol{z}|\boldsymbol{g}, \boldsymbol{\theta}^{(n-1)}, \boldsymbol{f}^{(n-1)} \\ \boldsymbol{\theta}^n & \sim & p(\boldsymbol{\theta}|\boldsymbol{g}, \boldsymbol{z}^{(n-1)}, \boldsymbol{f}^{(n-1)}) \end{array} \right. \tag{2}$$

One iteration of the Gibbs sampler

where $\boldsymbol{\theta}$ stands for the hyperparameters of the priors, i.e. the mean and variances of the gaussian laws $\mathcal{N}_1$ and $\mathcal{N}_2$, and n is the iteration index.

3 Bayesian segmentation in the wavelet domain

In order to reduce the computation time of the Gibbs sampler the initial bayesian segmentation based a Potts model has been projected in the wavelet domain [6]. It segments, in a bottom-up scheme, the scaling and wavelet subbands, by :
a) considering the wavelet coefficients as a mixture of only two gaussians, which leads to a segmentation in two classes.
b) incrementally segmenting all subbands from the initial coarse scaling coefficients segmentation, thus sparing segmentation iterations and time.
The decomposition of our observable $\tilde{g}$ in the wavelet domain can be expressed :

$$\tilde{g}(g, V_j, W_j) = P(g, V_J) + \sum_{j=1}^{J} P(g, W_j) \tag{3}$$

Observable decomposed in the wavelet domain

with V_j and W_j respectively the scaling and wavelet subbands, and $j = \{0...J\}$ the scale (J is the coarsest scale).
In order to operate a Bayesian segmentation in the wavelet domain, we recall that a new PMRF model was used in [6]. This one takes into account the privileged orientations of the coefficients in an orthogonal wavelet decomposition, i.e. vertical, diagonal and horizontal. It thus considers a second order neighborhood for the PMRF, which leads to the new expression of this PMRF :

$$\begin{aligned} p(z(i,j),(i,j)\in\mathcal{R}) = \tfrac{1}{T(\alpha_V,\alpha_{D_1},\alpha_{D_2},\alpha_H)} \times \exp\Big\{ & +\alpha_V \sum_{(i,j)\in\mathcal{R}} \delta(z(i,j)-z(i-1,j)) \\ & +\alpha_{D_1} \sum_{(i,j)\in\mathcal{R}} \delta(z(i,j)-z(i+1,j-1)) \\ & +\alpha_{D_2} \sum_{(i,j)\in\mathcal{R}} \delta(z(i,j)-z(i-1,j-1)) \\ & +\alpha_H \sum_{(i,j)\in\mathcal{R}} \delta(z(i,j)-z(i,j-1)) \Big\} \end{aligned} \tag{4}$$

PMRF tuned to the privileged orientations of the wavelets subbands (second-order neighborhood)

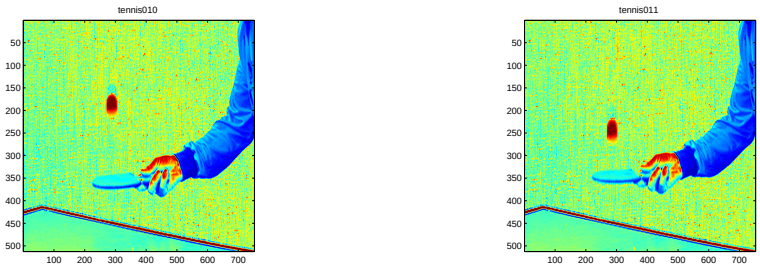

Figure 1: Original frames 10 and 11 of the "tennis" sequence, upsampled here to 752×512 for the needs of the segmentation and of the ME tests.

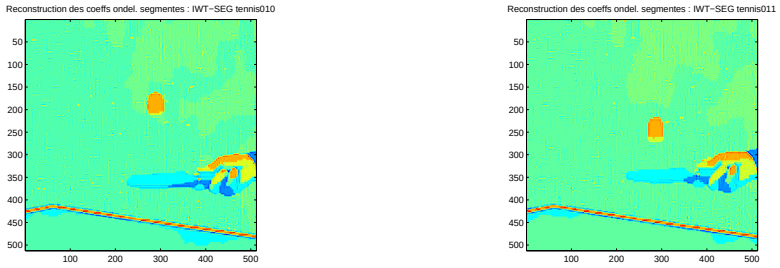

Figure 2: a) Final unsupervised segmentation, with a two-scales decomposition and six classes, of frames 10 and 11.

4 Hierarchical segmentation

A first innovation in our new hierarchical segmentation-ME scheme, is to make the hypothesis that there is a close correlation between all the corresponding subbands of two successive frames. This observation leads us to segment now the coarse scaling subband of any image $n+1$ from the result of the segmentation of the corresponding subband of image n. We then can improve the segmentation scheme by using the segmentation of scaling coefficients of image I_n to initialize the segmentation of the scaling coefficients of image I_{n+1}. This can be very interesting for large frames where the initial image segmentation is long.
A second innovation is that we now reconstruct the image at each level of the decomposition and from the segmented subbands at each level. This is a trivial step which brings a set of 2L segmented frames. The same rules are applied for the filtering, i.e. the scaling subband regions are replaced by the a "region averaging" of the original scaling coefficients. The wavelets subband regions, for $K = 2$ are replaced by their original wavelet coefficients and the coefficients, for $K = 1$, are zeroed.

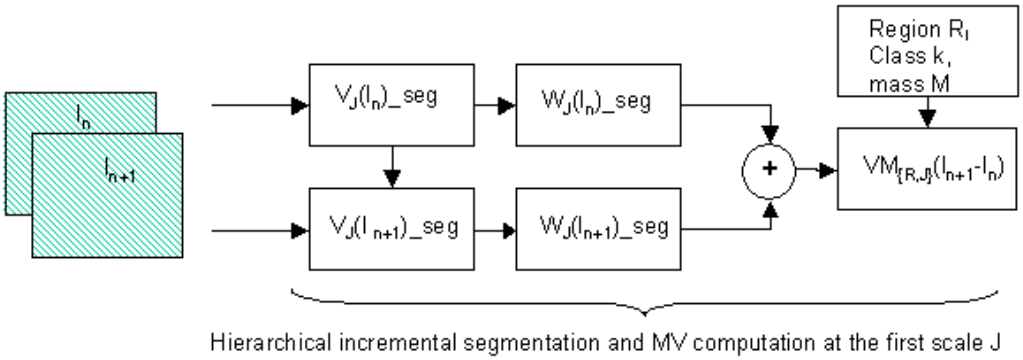

Figure 3: Detailed part of the incremental segmentation from two successive frames at scale $j = 2$, and computation of the MV at this same scale. This step has been added to the wavelet segmentation step and is done at each scale j for two successive frames, after the segmentation of their subbands at the same scale j and after a reconstruction from the segmented subbands.

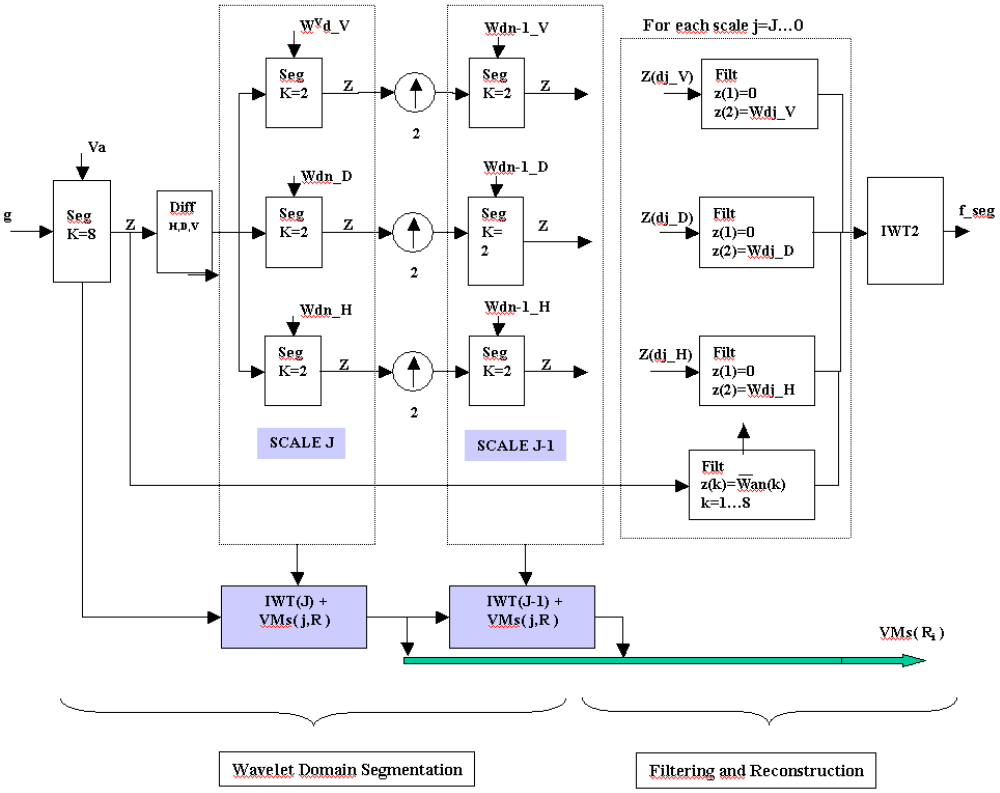

Figure 4: Wavelet domain segmentation scheme with the new reconstruction and ME computations at each scale

5 Hierarchical motion estimation

Based on the region-based segmentation, we can compute now the ME of regions or objects at each level. This approach offers two possibilities :
1) to compute the MVs at a coarse scale, thus offering the property of scalability which is interesting (see MPEG4 standard) for the computation speed and the adaptativity to the complexity of the decoder.
2) the ability to make an inter-scale correlation of the MVs, from the coarse to the fine scale, and thus to increase the robustness of these MVs. An illustrative example would be the "Edberg" sequence where the tennis player is a global "object", with its global motion, which is itself composed of several smaller objects (the members) having their own local motion. In such a case the low resolution analysis can easily provide a global motion which is not the case if we compute the VM at a high resolution.

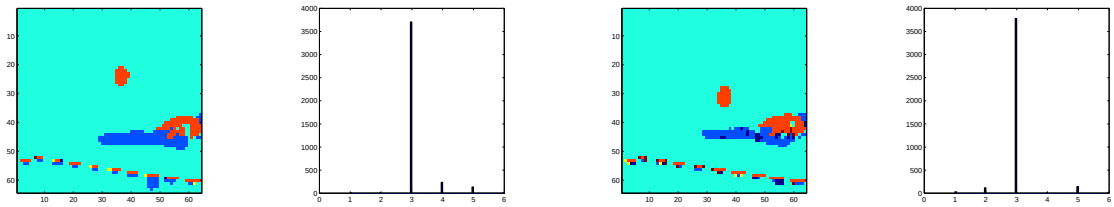

Figure 5: Incremental segmentation from the coarsest scale ($j = J = 3$) of two successive frames, 10 and 11. The histograms are shown to check that, at this coarsest scale, we start the segmentation with the requested number of classes, i.e. $K = 6$. This coarse resolution segmentation highlights the fact that there are four main regions in the frame : the sprite (almost static) background, the ball, the hand and the racket. The coarse subbands is thus able, better than other scales, to give a simple, condensed, representation of the scene by a "key" segmentation. This key segmentation is a the basis of the ME scheme robustness.

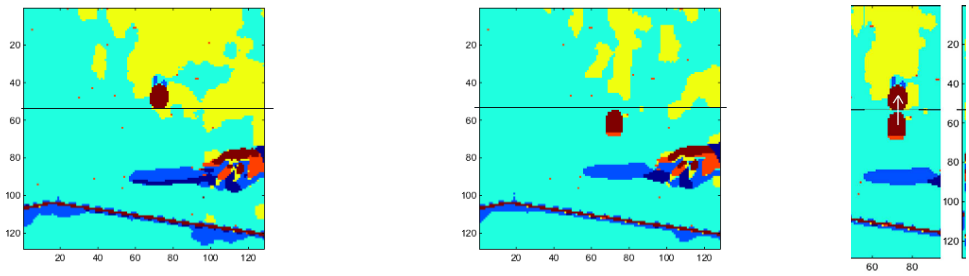

Figure 6: a) Reconstruction after segmentation at scale $j = 2$ for frame 10 b) same for frame 11 c) MVs computation based on the motion of the mass center for the "ball" region. The tracking of the ball between frames 10 and 11 is based on a "mass correspondance" principle [5], i.e. a close number of pixels pertaining to the same class k, and in a fixed neighborhood of the ball, between these frames.

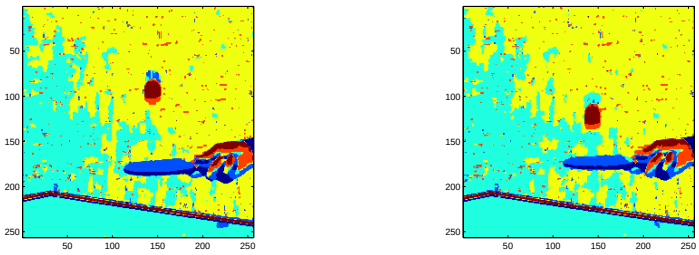

Figure 7: a) Reconstructed image of segmentation at scale $j = 1$ on frame 10 b) same for frame 11. MVs between scales have a strong correlation which is obvious on this higher resolution frame. We thus are able to adopt the same reasoning for MVs than for the spatial segmentation : an initialization by the MV of the former scale must induce a faster computation of the MV at the present scale on the same object. The use of a cross-correlation distance (region-based matching) rather than a difference between masses (see e.g. : the search of an optimal MV in an energy based ME [11]) could bring more robustness in our scheme.

6 Conclusion

In this work we have shown a new approach of the motion estimation in a region-based (or object-based) framework and in a hierarchical scheme. The originality of this scheme resides in the hierarchical computation of MVs on regions, together with the ability to transmit these MVs at each scale, which enables the scalability of their transmission. This scheme is based on the combination of a wavelet multiresolution, of a bayesian spatial segmentation performed in the wavelet domain, on the acceleration of the segmentation for correlated images (incremental segmentation scheme) and of the computation of MVs on the displacement of the mass center of regions (or objects). The improvement of region to object can be made, if we have no initial knowledge of the regions tracked, by using a shape recognition algorithm [4].

References

[1] Andersson, K. and Knutsson, H., Multiple Hierarchical Motion Estimation, Proceedings of the IASTED International Conference on Signal Processing, Pattern Recognition and Applications, pp. 80–85, June 25-28, 2002.

[2] Andrieu, C., Doucet, A. and Duvaut, P., Methodes de Monte-Carlo par Chaines de Markov appliquees au traitement du signal, rapport interne ETIS/ENSEA-URA/CNRS 2235 97-N 3, pp 1–27, 2003.

[3] Amit, Y., Geman, D. and Fan, X., A coarse-to-fine strategy for multi-class shape detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, V. 28, pp 1606–1621, 2006.

[4] Brault P. and Mounier H., Automated Transformation-Invariant Shape Recognition Through Wavelet Multiresolution, SPIE01 International Society for Optical Engineering , San Diego, USA, 2001.

[5] Brault P. and Mohammad-Djafari A., Bayesian Segmentation of Video Sequences Using a Markov-Potts Model, WSEAS Transactions on Mathematics, Vol.3 (1), pp 276–282, 2004.

[6] Brault P. and Mohammad-Djafari A., Unsupervised Bayesian Wavelet Domain Segmentation Using a Potts-Markov Random Field Modeling, Journal of Electronic Imaging, Vol 14(4), 2005.

[7] Brault, P., Motion Estimation and Segmentation ; I) Motion-tuned wavelets for ME II) Bayesian unsupervised segmentation of sequences in the wavelet domain, PhD thesis, Universite Paris-Sud, France, november 2005.

[8] Demoment, G., Giovannelli, J.F. and Mohammad-Djafari, A., Laboratoire des Signaux et Systèmes, Problemes inverses en traitement du signal et de l'image, Techniques de l'Ingenieur, Traite Telecoms, TE 5 235, pp 1–31, Novembre 2001.

[9] Dugelay J.L. and Sanson H., Differential methods for the identification of 2D and 3D motion models in image sequences. EURASIP image communication, 7:105–127, 1995.

[10] Féron O. and Mohammad-Djafari A., Image fusion and unsupervised joint segmentation using a HMM and MCMC algorithms, J. of Electronic Imaging, vol. 14(2), paper n023014, April 2005.

[11] Jodoin P.M. and Mignotte M., Markovian segmentation and parameter estimation on graphics hardware, J. of Electronic Imaging, to appear in 2006.

[12] Gangaputra S. and Geman D., The Trace Model for Object Detection and Tracking, to appear in lecture notes on Computer Science, 2006.

[13] Gangaputra S. and Geman D., A Unified Stochastic Model for Detecting and Tracking Faces, Proceedings of the second canadian conference on computer and robot vision (CRV05), pp 306–313, 2005.

[14] Krempp, S., Geman, D and Amit, Y., Sequential Learning of Reusable Parts for Object Detection, Technical Report, John Hopkins University, Baltimore, Maryland, 2002.

[15] Qian G. and Chellappa R., Bayesian Algorithms for Simultaneous Structure from Motion Estimation of Multiple Independently Moving Objects, IEEE Transactions on Im. Proc., vol. 14, Jan. 2005.

[16] Richardson I., H264 and MPEG4 Video Compression, Wiley, 2003.

[17] P. Salembier, F. Marqués, M. Pardàs, R. Morros, I. Corset, S. Jeannin, L. Bouchard, F. Meyer, and B. Marcotegui, Segmentation-based video coding system allowing the manipulation of objects. IEEE Trans. on Circuits and Systems for Video Technology, 7(1):60-74, February 1997.

[18] Tziritas, G. and Labit, C., Motion Analysis for Image Sequence Coding, Elsevier Science, J. Biemond eds., Delft University of Technology, The Netherlands, 1994.