Motion Vector Estimation and Encoding for Motion Compensated DWT

Marco Cagnazzo[1,2], Valéry Valentin[1], Marc Antonini[1], Michel Barlaud[1]

[1] I3S Laboratory, UMR 6070 of CNRS, University of Nice-Sophia Antipolis
Bât. Algorithmes/Euclide, 2000 route des Lucioles - BP 121 - 06903 Sophia-Antipolis
Cedex, France. Phone: +33(0)4.92.94.27.21 — Fax: +33(0)4.92.94.28.98
{vvalenti,am,barlaud}@i3s.unice.fr
[2] Dipartimento di Ingegneria Elettronica e delle Telecomunicazioni,
Università Federico II di Napoli, via Claudio 21 - 80125 Napoli, Italy
cagnazzo@unina.it

Abstract. In this work, we propose a new technique for estimation and encoding of motion vectors, in order to achieve an efficient and scalable representation of motion information. The framework is Motion Compensated Three-Dimensional Wavelet Transform (MC3DWT) video coding. At low bit-rates an efficient estimation and encoding of motion information is especially critical, as the scarce coding resources have to be carefully shared between motion vectors and transform coefficients. The proposed technique, called Constrained Motion Estimation, outperforms the usual "unconstrained" one at low to medium rates, and is essentially equivalent to it at higher rates.
Moreover, the proposed encoding technique for Motion Vectors, based on Wavelet Transform and context-based bit-plane coder, gives a scalable representation of them.

1 Introduction

The ability of exchange, store and transmit information, and most often *visual* information like graphs, images, videos, has become more and more important in recent years, not only for enterprises or academic institutions, but also for common people. On the other hand, this kind of information requires many resources for storage and transmission. It is not surprising, then, the huge effort that has recently been addressed to image and video compression issues, and that has allowed the deployment of successful international standards, like JPEG and JPEG-2000, the MPEG and H.26x series [1, 2].

Yet, even the most recent video standards suffer from some problems, like artifacts from block-based transform techniques, a not completely satisfying support to scalability, and a suboptimal bit-rate allocation between motion information and residual coding. In particular, this problem is especially important at very low and low bit rates, where a full representation of motion information can easily take up an unfairly large part of coding resources.

In order to overcome these problems, subband coding and namely Wavelet Transform (WT) based techniques have been often proposed as an alternative framework (with respect to the standardized DCT-based hybrid approach) both for image and video coding. In particular, for video coding, three dimensional video coding techniques have been studied for several years [3]. Even in this framework, the importance of Motion Estimation (ME) and Compensation (MC) has been early recognized [4], giving rise to the so called Motion Compensated 3DWT techniques, which currently are among the most promising approaches to the video coding issue, as they have competitive performances and provide natural support to scalability [5, 6].

The scalability issue is of increasing importance as heterogeneity seems to be one of the most persisting features of networks in general and of the Internet in particular. In other words, users with different resources (in terms of both bandwidth and computing power) want different quality for multimedia contents, and in order to accomplish all requests without encoding many times the original data, a *scalable* encoding algorithm has to be employed. This demand has been acknowledged in the new still image standard JPEG2000, but it has not been completely satisfied in video standards.

In this paper we propose a technique for motion vectors estimation and encoding, in the framework of MC3DWT video coding, as described in [5, 7]. The encoder consists of a Temporal Stage, in which temporal filtering is performed via a Motion Compensated Lifting Scheme, and a Spatial Stage, in which bidimensional WT is carried out. In this work we focus on the Temporal Stage. The new ME algorithm, described in section 2, reduces the rate needed for motion vectors, while the encoding technique, which is discussed in section 3 gives an efficient and scalable representation of them. Finally, conclusions are reported in section 4.

2 Motion Estimation Technique

2.1 Motion Compensated Lifting Scheme

In this work, Motion Compensated Lifting Scheme has been used to perform the WT along temporal axis. More generally, the lifting scheme is used as efficient implementation of and, as shown in [8], a wavelet filter bank can be implemented by a lifting scheme. Let us see, for example, how this is possible for the biorthogonal 5/3 filter. Let I_t be the t^{th} frame, and h_t (resp. l_t) the t^{th} high (resp. low) subband obtained after WT. The 5/3 filter can undergo the following decomposition:

$$\begin{cases} h_t(x,y) = I_{2t+1}(x,y) - \frac{1}{2}[I_{2t}(x,y) - I_{2t+2}(x,y)] \\ l_t(x,y) = \quad I_{2t}(x,y) + \frac{1}{4}[h_{t-1}(x,y) + h_t(x,y)] \end{cases} \tag{1}$$

This is also known as $(2, 2)$ lifting scheme. In (1), motion was not taken into account, but, in a typical video sequence, there is a lot of movement due to both camera panning and zooming, and objects displacement and deformation. Then, if we simply use (1) on a video sequence, we end up to apply this transform to a signal characterized by many sudden changes, that is, hard to compress.

Motion Compensation can be successfully introduced in order to overcome this problem. Here we use MC as described in [5] and [9], that is, with the basic idea of carrying out the temporal transform along motion direction. Let us suppose that we know the backward and forward MVFs, indicated as $BW_t(x,y)$ and $FW_t(x,y)$, and representing the position that pixel (x,y) of the frame t has, respectively, in frame $t-1$ and $t+1$. Then we can modify (1) as follows:

$$\begin{cases} h_t(x,y) = I_{2t+1}(x,y) - \frac{1}{2}\left[I_{2t}(BW_{2t+1}(x,y)) - I_{2t+2}(FW_{2t+1}(x,y))\right] \\ l_t(x,y) = \quad I_{2t}(x,y) + \frac{1}{4}\left[h_{t-1}(BW_{2t}(x,y)) + h_t(FW_{2t}(x,y))\right] \end{cases} \tag{2}$$

This Motion Compensated Lifting Scheme is perfectly invertible and is widely used in literature [10–12].

If movement is accurately estimated, motion compensated wavelet transform will generate high frequency bands with low energy, and a low frequency band in which objects position is precise and their shape is clear. Thus, motion compensation allows preserving both high coding gain and temporal scalability. Note that, as we use the (2,2) lifting scheme, only two MVF, a backward and a forward one, are needed per each frame.

2.2 Unconstrained Motion Estimation

In the context of block-based motion estimation algorithm, a block $B_t^{(\mathbf{p})}$ from frame t, centered on pixel $\mathbf{p}$ is compared to a block $B_\tau^{(\mathbf{p}+\mathbf{v})}$, that belongs to a reference frame (for example, the previous one, so $\tau = t-1$) and that is no longer centered in $\mathbf{p}$, but is displaced by a vector $\mathbf{v}$. The estimated motion vector $\mathbf{v}^*$ is chosen as the one minimizing some metric (i.e. a measure of error) d between the two blocks:

$$\mathbf{v}^* = \arg\min_{\mathbf{v}\in W}\left[d\left(B_t^{(\mathbf{p})}, B_\tau^{(\mathbf{p}+\mathbf{v})}\right)\right] \tag{3}$$

where W is the allowed search window for block matching. Various metrics have been proposed to play the role of $d(B_1, B_2)$, the MSE being the most popular. In [9] the ZNSSD (Zero-mean Normalized Sum of Squared Differences) was proposed, as it provides a ME independent from frame-to-frame average intensity variation.

The ME criterion described in (3) can be used to find both BW and FW MVFs, by suitably choosing a value for τ.

2.3 Constrained Motion Estimation

In Motion Compensated video coding, we have to split up the total available bit-rate between MVFs and motion compensated coefficients, being aware that, giving more resources to MVFs yields better ME and thus reduces the energy of high frequency bands, but, on the other hand, reduces also the bit-rate available to encode them. When the bit-rate is low or very low, the motion information can easily grow up to constitute a remarkable share of the total coding resources, so efficient or possibly *lossy* coding techniques become of great interest. In this section we propose a first solution to this problem, trying to decrease the bitrate needed by MVS without degrading too much motion information just by regularization.

Here we propose a ME technique that gives a substantial reduction of the rate needed for MVF's encoding. The basic idea is to impose some reasonable constraints to the ME criterion, in order to get a MVF that can be efficiently encoded and that anyway remains a good estimation of motion. In [9] we proposed a simple simmetry constraint: we want the backward MVF to be the opposite of the forward one. Then we searched for a displacement vector, which under this constrain minimizes a quantity depending on both the forward and the backward error. In that work we chose the sum of them:

$$\mathbf{v}^* = \arg\min_{\mathbf{v}\in W} J(\mathbf{v}) \tag{4}$$

where W is a suitable search window and

$$J(\mathbf{v}) = \left[d\left(B_t^{(\mathbf{p})}, B_{t-1}^{(\mathbf{p}-\mathbf{v})}\right) + d\left(B_t^{(\mathbf{p})}, B_{t+1}^{(\mathbf{p}+\mathbf{v})}\right)\right] \tag{5}$$

This criterion allowed us to obtain a good estimation of motion with a reduced encoding cost, as the constraint makes it possible to encode just one of the MVFs (the other is deduced by symmetry).

Here we modify the criterion (4), with the aim of obtaining MVFs even more efficiently encodable. Indeed, even though the symmetry constraint implies some regularity, the MVFs can still suffer from some bad estimation which causes irregularities: see for example Fig.1, where it is reported an estimated MVF for the "foreman" sequence: in the quite homogeneus helmet area, the MVF, even if minimizes the metric, has a remarkable entropy.

Hence we introduce some other constraints: a "length penalty" and a "spatial variation penalty". The new criterion is expressed as follows:

$$\mathbf{v}^* = \arg\min_{\mathbf{v}\in W} \left[J(\mathbf{v}) + \alpha\frac{||\mathbf{v}||^2}{||\mathbf{v}||^2_{max}} + \beta(||\nabla v_x||^2 + ||\nabla v_y||^2)\right] \tag{6}$$

In Fig.2 the regularized MVF is shown. It is clear that the constraints help in regularizing the MVF. Initial experiments showed the existence of values for α and β which allow a fair regularization without degrading too much the motion information. We can gain a deeper sight on this phenomenon by evaluating the effect of regularization on the first order entropy of regularized MVFs and the respective prediction MSE, see Tab.1. These results were obtained for the test sequence "foreman", with a block size of 16×16 and whole pixel precision.

In this table we also reported the entropy of Wavelet Transformed (3 level dyadic decomposition) MVFs, in order to show that WT allows reducing entropy, and that regularization is effective even in the wavelet domain. This results suggest us to look for an encoding technique that makes use of WT of regularized MVFs, like described in section 3. We also remark that with a suitable choice of parameters, we can achieve an entropy reduction of 16% (and even 40% in WT domain), while the prediction MSE increases only of 1.5%.

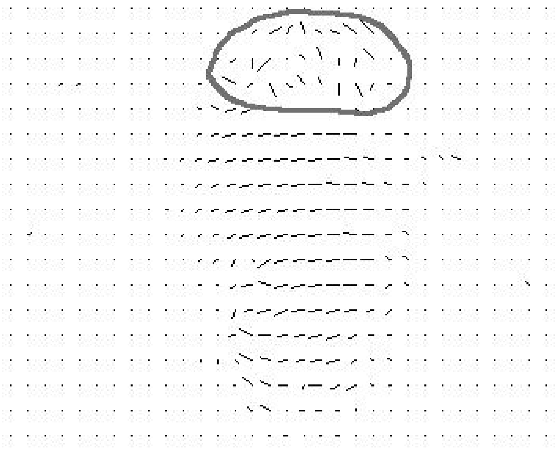

Fig. 1. Irregular motion vector estimation

Other resultes for the test sequence "foreman", are presented in Fig. 3. Here we compare the first order entropy of dense, whole pixel-precision MVFs (regularized and non regularized), when different levels of dyadic WT are applied. Of course, 0 levels stands for no WT. The MVF has whole pixel precision, with varying regularization parameters. We remark that wavelet transform is indeed more efficient on regularized dense MVF: it significantly reduces the entropy of MVFs suggesting for a successful application of a WT based encoding technique.

Besides, it is useful to evaluate the impact of constrained MVFs on the whole video coder. So, in Fig.4 we compared global RD performances of our MC3DWT video coder when the usual unconstrained ME and the proposed constrained ME criterion are used. This results were obtained on the "foreman" sequence, with regularization parameters $\alpha = 10$ and $\beta = 5$, precision of whole pixel, and block size of 16×16. Again, the rates for MVFs were assessed by computing their entropy. The graph shows that the proposed method yields globally better performances, especially at low to medium rates, where we achieve up to 1.3 dB of improvement with respect to the usual unconstrained technique. This result confirm the intuitive idea that at low rates it is better to have an approximate but cheap description (in term of needed encoding resources) of motion and to dedicate more resources to transform coefficients.

3 Scalable Motion Vector Encoding by Wavelet Transform

As already pointed out, an efficient representation of motion information can not be the same both at high and at low bit-rates: when the resources are meager, it becomes interesting to dispose of a *lossy* encoding technique for MVFs. Moreover, if we think about the heterogeneity of networks and users, scalability

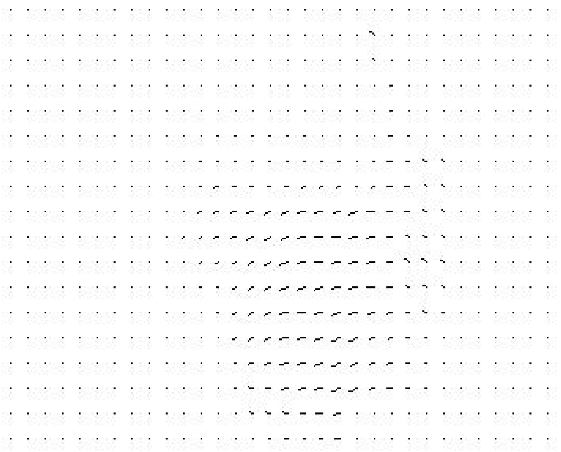

Fig. 2. Regularized motion vector field

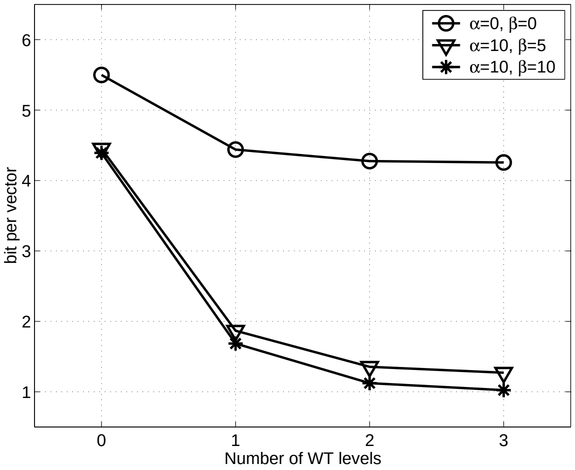

Fig. 3. Effect of WT on regularized MVF Entropy

also assumes an increasing importance. Hence we propose an *embedded* encoding algorithm, which then assures low cost encoding when scarce encoding resources are available, and the ability of lossless encoding when we dispose of a high bit-rate.

3.1 Technique Description

Let us see how the proposed technique accomplishes these demands. First of all we compute a dense MVFs (that is, a vector for each pixel) at high precision

α	β	Entropy of MVF [bit/vector]	Entropy of WT [bit/vector]	Prediction MSE
0	0	4.3322	0.55684	48.2169
0	4	3.7249	0.3453	49.902
0	20	3.3672	0.258	56.5364
10	0	3.9328	0.46946	48.2395
10	4	3.6126	0.33895	49.9652
10	20	3.3019	0.25366	56.4439
30	0	3.8094	0.44627	48.3522
30	4	3.5822	0.33686	50.077
30	20	3.2822	0.25074	56.5431
100	0	3.6244	0.41705	48.9757
100	4	3.4587	0.33033	50.5816
100	20	3.1958	0.24644	56.912

Table 1. Regularization parameters effect – test sequence “foreman”, ME with 16×16 blocks and whole pixel precision

(i.e. quarter pixel or better), obtained by B-Spline interpolation [13]; indeed, we are going to encode MVFs with a scalable technique allowing both lossless and lossy reconstruction, so we leave to the MVF encoder the job of rate reduction, while in the ME we simply get the most complete information about movement.

Then, vertical and horizontal components of MVFs undergo a JPEG-2000 like compression scheme: bidimensional WT is performed on them, with a decomposition structure that can vary, but that we initially chose as the usual dyadic one. We can use both integer filter like the (5/3),which alllows perfect reconstruction even with finite precision arithmetics, and non integer – but better performing – like the (9/7). They are implemented by lifting scheme. First experiments suggested us to use three decomposition levels on each component of motion vectors. The resulting subbands are encoded with a context-based bitplane encoder, which gives an efficient and scalable representation of MVFs: by increasing the number of decoded layers, we get an ever better MVF, and, when all layers are used, we get a lossless reconstruction, thanks to the use of integer filters.

3.2 Proposed Technique Main Features

The proposed technique aims to reproduce and generalize the behavior of Variable Size Block Matching (VSBM): in fact, lossy compression of WT coefficients tends to discard data from homogeneous area in High Frequency Subbands, and to preserve high activity areas: this is conceptually equivalent to adapt the resolution of motion information representation to its spatial variability, that is, to increase or decrease the block size like in VSBM, but with the advantage of a greatly extended flexibility in representation of uniform and non uniform areas, which are no longer constrained to rectangular or quad-tree-like geometry. This

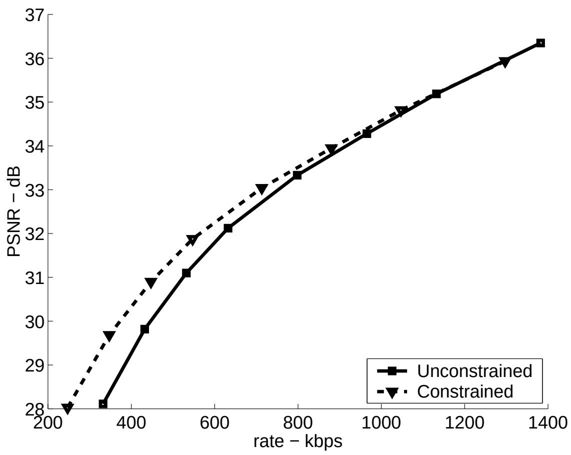

Fig. 4. Impact of ME methods on codec performances

is shown in Fig. 5 and 6, where a dense regularized ($\alpha = 10, \beta = 5$) MVF and the lossly compressed version are shown. The value of each component is represented in grayscale, with medium gray standing for null component.

Another advantage of the proposed technique should be that it supplies a way to gracefully degrade motion information, so that it becomes easier to find the best allocation of the total available rate R_T between MVFs R_{mv} and coefficients R_{sb}. In fact it is clear that it must exist an optimal split up: performances usually improve from the case we do not use MC ($R_{mv} = 0$) to the one in which it is used ($R_{mv} > 0$), and it is also clear that performances decrease when R_{mv} tends to saturate R_T. With the proposed technique we are able to smoothly vary the rate addressed to MVFs, and so we can more easily find the optimal allocation.

Some other results of this coding technique in the lossy case are also shown in Fig. 7: here we reported the Prediction MSE of lossly encoded MVFs in function of the coding rate, when the filter used for WT is changed. We note that it is possible to strongly reduce the rate without increasing too much the MSE. We also remark that the 9/7 filter has better performances in a wide range of rates.

In [14] scalability of MVF was obtained by undersampling and approximating the estimated vectors, an by refining them in the enhancement layers, and the scalability was strictly related to subband decomposition. Here, instead, the proposed algorithm ensures a more flexible embedded description of MVFs.

4 Conclusions

In this work a new technique for estimation and encoding of MVF is proposed. It gives a compact and scalable representation of MVFs, which is important, above all at very low to low bit-rates, in order to achieve an efficient and fully

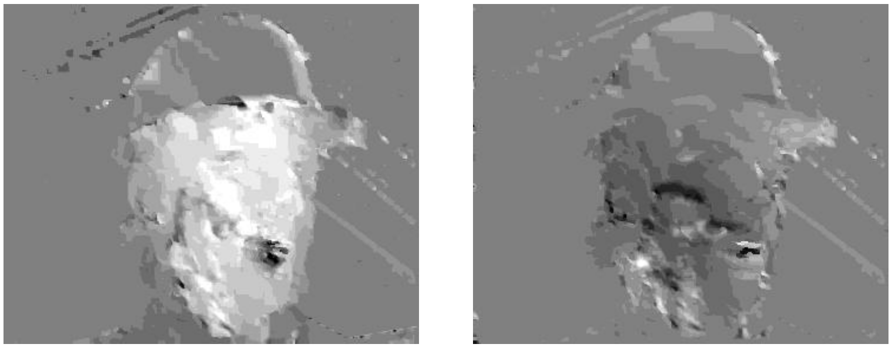

Fig. 5. Original dense MVF, frame 5, horizontal and vertical component. Entropy > 10 Mbps, rate 3.5 Mbps with lossless compression

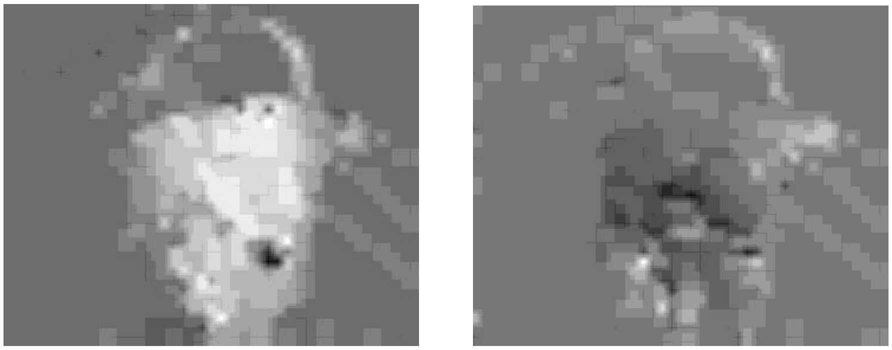

Fig. 6. Decoded MVF, frame 5, horizontal and vertical component. Encoding rate 0.02 bit/vector, or 75 kbps with lossy compression

scalable encoded bitstream. Moreover, an efficient representation of MVFs is needed in order to achieve the optimal trade-off between motion information and subbands. In these work we showed the existence of such an optimal trade off, and in future we going to deeply analyze the MVF encoding technique which allows us to gracefully degrade MVFs in order to find the optimal rate allocation.

References

1. ISO/IEC JTC1/SC29/WG11: ISO MPEG4 Standard. (1998)
2. Joint Video Team of ISO/IEC MPEG and ITU-T VCEG: Joint Committee Draft, JVT-C167. (2002)
3. Karlsson, G., Vetterli, M.: Three-dimensionnal subband coding of video. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Volume 2., New York, USA (1988) 1100–1103
4. J.-R. Ohm: Three dimensional subband coding with motion compensation. IEEE Transactions on Image Processing **3** (1994) 559–571

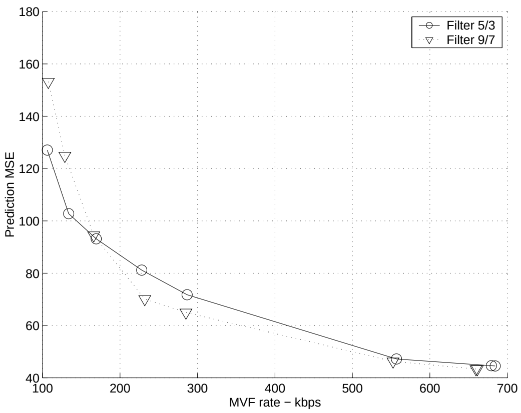

Fig. 7. Scalable Lossy coding of MVFs

5. Parisot, C., Antonini, M., Barlaud, M.: Motion-compensated scan based wavelet transform for video coding. In: Proceedings of Tyrrhenian International Workshop on Digital Communications, Capri, Italy (2002)
6. Viéron, J., Guillemot, C., Pateux, S.: Motion compensated 2D+t wavelet analysis for low rate fgs video compression. In: Proceedings of Tyrrhenian International Workshop on Digital Communications, Capri, Italy (2002)
7. Parisot, C., Antonini, M., Barlaud, M.: 3d scan-based wavelet transform and quality control for video coding. EURASIP Journal on Applied Signal Processing (2003) 56–65
8. Daubechies, I., Sweldens, W.: Factoring wavelet transforms into lifting steps. J. Fourier Anal. Appl. **4** (1998) 245–267
9. Valentin, V., Cagnazzo, M., Antonini, M., Barlaud, M.: Scalable context-based motion vector coding for video compression. In: Proceedings of Picture Coding Symposium, Saint-Malo, France (2003) 63–68
10. Secker, A., Taubman, D.: Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting. In: Proceedings of IEEE International Conference on Image Processing, Thessaloniki, Greece (2001) 1029–1032
11. Flierl, M., Girod, B.: Investigation of motion-compensated lifted wavelet transforms. In: Proceedings of Picture Coding Symposium, Saint-Malo, France (2003) 59–62
12. Tillier, C., Pesquet-Popescu, B., Zhan, Y., Heijmans, H.: Scalable video compression with temporal lifting using 5/3 filters. In: Proceedings of Picture Coding Symposium, Saint-Malo, France (2003) 55–58
13. Unser, M.: Splines: A perfect fit for signal and image processing. IEEE Signal Processing Magazine **16** (1999) 22–38
14. Bottreau, V., Bénetière, M., Felts, B., Pesquet-Popescu, B.: A fully scalable 3d subband video codec. In: Proceedings of IEEE International Conference on Image Processing, Thessaloniki, Greece (2001) 1017–1020