4D Scalable Multi-View Video Coding Using Disparity Compensated View Filtering and Motion Compensated Temporal Filtering

Jens-Uwe Garbas, Ulrich Fecker, Tobias Tröger and André Kaup
Chair of Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany
Email: {garbas, fecker, troeger, kaup}@LNT.de

***Abstract*— In this paper, a novel framework for scalable multi-view video coding is described. A well known wavelet based scalable coding scheme for single-view video sequences has been adopted and extended to match the specific needs of scalable multi-view video coding. Motion compensated temporal filtering (MCTF) is applied to each video sequence of each camera. The use of a wavelet lifting structure guarantees perfect invertibility of this step, and as a consequence of its open-loop architecture, SNR and temporal scalability are attained. Correlations between the temporal subbands of adjacent cameras are reduced by a novel disparity compensated view filtering (DCVF), method which is also lifting based and open-loop to enable view scalability. Spatial scalability and entropy coding are achieved by the JPEG2000 spatial wavelet transform and EBCOT coding, respectively. Rate allocation along the temporal-view-filtered subbands is done by means of an RD-optimal algorithm. Experimental results show the high scaling capability in terms of SNR, temporal and view scalability.**

I. Introduction

Multi-view video coding (MVC) is one of the most interesting technologies for future multimedia applications such as free-viewpoint video or photorealistic rendering of 3D scenes [1]. Unfortunately, simultaneous recording of a moving scene with several cameras generates huge amounts of data which makes efficient compression necessary. The most straightforward solution to MVC is to independently code each video of each camera with a conventional video codec, e.g. H.264/AVC [2]. As the efficiency of this so-called simulcast coding is not optimal, much effort has been spent to find better solutions. In [3], it could be shown that there exists significant correlation across the different views at the same or similar time instances. In order to achieve efficient compression results, exploitation of these inter-view dependencies is indispensable. Most multi-view codecs that have been published recently use some kind of inter-view prediction and show superior performance compared to simulcast coding. The new MVC reference software used for MPEG evaluations [4] is an example for this strategy.

With the increasing diversity of applications and network environments, another technology has also been in the focus of current research activities: scalability. It has become a very important feature for video coding and modern communication scenarios to serve different devices as well as to adapt to bandwidth variations. To this end, most scalable video codecs (SVC) such as [5] provide spatio-temporal-SNR scalability. Image and video coders based on wavelet transforms have proven to be an excellent way to naturally achieve such scalability features. In 1994, the so-called motion compensated temporal filtering (MCTF) was first introduced in [6] and further developed in [7]. This subband filtering operation along the motion trajectories of the temporal axis of a video sequence inherently provides temporal scalability and allows easy integration of SNR scalability because of its open-loop structure. Spatial wavelet transform and embedded coding of the temporal subbands, for example with JPEG2000 as in [8], finally leads to a highly scalable representation of the video.

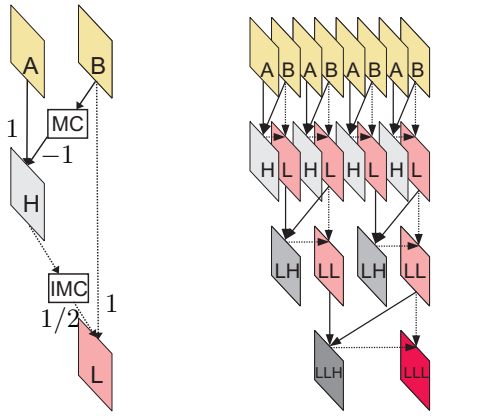
In this paper, we propose a combination of the two above-mentioned technologies: Exploitation of inter-view dependencies of multi-view sequences in a highly scalable framework. View scalability is achieved by extending a well-known 3D scalable coding scheme by a 1D wavelet transform along the view axis, which we call disparity compensated view filtering (DCVF). As far as we know, a wavelet approach to MVC has only been reported in [11]. In [11], a view-directional filtering similar to ours is proposed, but only applied to temporal lowpass bands. Scalability is not analyzed in this paper at all. Therefore, in our paper the first approach to scalable MVC based on a complete 4D wavelet transform is described. Furthermore, our scheme is compatible with JPEG2000.

II. Proposed Method

We first describe the principles of motion compensated temporal filtering in Sec. II-A. The new disparity compensated view filtering is introduced in Sec. II-B, followed by a method for spatial transform and coefficient coding with optimal rate allocation via JPEG2000 in Sec. II-C.

A. Motion Compensated Temporal Filtering

Motion compensation (MC) is a key technique of most video coding schemes to reduce the temporal correlation. In MCTF based codecs, MC is used to apply a multiresolution subband transform along the motion trajectories. Thus, after the transform, the temporal lowpass-band concentrates more energy while the highpass bands are more close to zero than without motion alignment. However, since this is a non-linear process, the direct integration of MC into a conventional wavelet transform prevents invertibility. Fortunately, any pair



(a) (b)

Fig. 1. MCTF with Haar Filter. (a) Transform step $A/B \rightarrow L/H$ in lifting structure (b) Multiresolution wavelet tree generated by cascaded transform steps

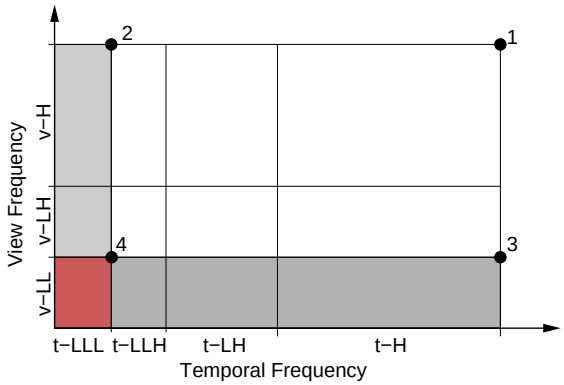


Fig. 2. View-temporal subband decomposition structure of multi-view video data

of biorthogonal filters can be implemented in a lifting structure [9]. In [8] and [10], it was shown that this guarantees perfect reconstruction during synthesis in any case, regardless of the MC.

A lifting implementation consists of three basic steps: split, predict, update. In the first step, the video sequence is split into two subsets, say A and B. In the predict step, the (even) A-Frames are predicted by the (odd) B-Frames via a prediction filter $P(z)$ and the highpass wavelet coefficients (H-Frames) are calculated as the prediction error. In the update step, the the lowpass coefficents (L-Frames) are generated as sum of H-Frames filtered with an update filter $U(z)$ and the corresponding B-Frames.

The lifting implementation of the Haar-filter is shown in Fig. 1(a). In this simple case, the prediction and update filters are $P(\mathbf{z}) = -z_1^k z_2^l$ and $U(\mathbf{z}) = \frac{1}{2} z_1^{\bar{k}} z_2^{\bar{l}}$ respectively. The motion compensation (MC) and its inverse (IMC) are described by the k, l and $\bar{k}, \bar{l}$ shifts. This results in the motion compensated highpass and lowpass bands for the Haar transform:

$$\begin{aligned} H &= A - MC(B) \\ L &= 1/2 \cdot IMC(H) + B. \end{aligned} \quad (1)$$

The motion vectors (MVs) that are needed for MC are estimated by block matching troughout this paper. For the IMC, these MV-fields have to be inverted, which is in general not possible because there might be areas in the reference frame that are referred multiple times (e.g. areas becoming covered) and some that are not referred at all (e.g. uncovered areas). Nevertheless, we can simply invert the MVs for the unambiguous pixels in our coder. For the inversion of MVs that point to the same pixels multiple times, we simply choose one of them according to the scan order. The lifting structure guarantees invertibility of the transform even if the MC and IMC operations are not exactly inverses of each other.

If pairs of lowpass frames are further decomposed, subsequent levels of a wavelet tree are established, as can be seen exemplarily in Fig. 1(b). This MCTF with M decomposition levels produces $M + 1$ temporal subbands with most of the energy of the sequence compacted in the lowpass band. Provided that motion is well estimated and inverted in the update steps, the lowpass bands of each decomposition level can be used as temporal subsampled versions of the original video sequence, i.e. temporal scalability can easily be achieved by only partially inverting the MCTF.

After MCTF, a 2D spatial DWT would be applied to the temporal subbands in a wavelet based SVC, followed by embedded quantization and entropy coding. In order to extend this well-known 3D transform to the needs of MVC, we apply another 1D wavelet transform along the view axis after the MCTF and end up with a 4D transform. The view transform is described in detail in the next section.

B. Disparity Compensated View Filtering

As shown in [3], there is significant correlation along the view axis of a multi-view sequence, especially at the same time instances. This observation is still valid after MCTF of each view. To clarify this, one can think of the temporal lowpass subbands beeing a frame rate reduced version of the original video. If the structure of the MCTF is the same for all views, then the inter-view correlation of the L-frames must be nearly equal to the subsampled original frames. As the information in the highpass subband frames corresponds to the change of a scene over time and this change is similar for every view, there exists also significant inter-view correlation across the H-Frames.

For efficiently taking advantage of the view-directional correlation, disparity must be compensated. In multi-view sequences, there exists always global disparity since the different views are captured from different camera positions. As we also want to take into account local variations of this global disparity shift caused by several reasons, such as e.g. different camera angels and also artifacts caused by MCTF, we use block-based disparity compensation.

For the view transform, we can use exactly the same lifting

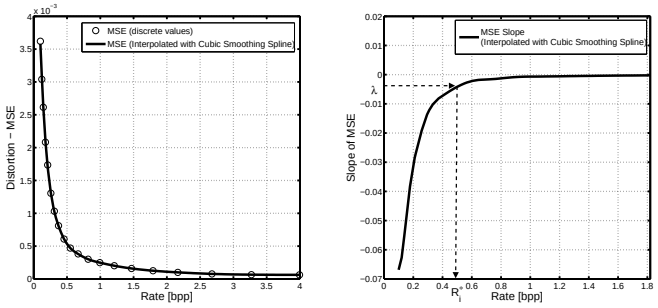


Fig. 3. Rate distortion curve (left) with spline approximation and its derivative (right) of t-LLL/v-LL subband

based filtering technique as for the temporal transform. We apply a so-called disparity compensated view filtering (DCVF) to each temporal subband. This view-directional transform reduces the inter-view correlation and is again inherently scalable. In Fig. 2, one can see the resulting subbands after MCTF (3 decomposition levels) and DCVF (2 decomposition levels). The points in the temporal-view-frequency plane denoted by 1 to 4 exemplify possible extraction points of this highly scalable scheme. Temporal and/or view scaled versions of the original multi-view sequence can be obtained by simply discarding all highpasss bands lying up and right of the specified point. For example, a multi-view-sequence with originally 8 views and a frame rate of 60 Hz can be completely reconstructed from point 1. Point 2 still contains the information of all cameras but with frame rate reduced by a factor of $2^3 = 8$ (7.5 Hz). Similary extraction of the subbands associated with 3 enables reconstruction at full frame rate but only every fourth (2^{-2}) video of the multi-view sequence is available. Finally, at point 4 we have only two views with a frame rate of 7.5 Hz left. This example demonstrates the high degree of scalability that is offered by the transform. Of course, with our framework other decomposition structures than in Fig. 2 are possible, for example different view-directional decomposition levels depending on the temporal frequency.

C. Spatial Wavelet Transform and Coding with JPEG2000

The resulting subbands after MCTF and DCVF are subjected to a 2D dyadic spatial wavelet transform and embedded block coding. This task is accomplished by the standard compliant implementation of JPEG2000 still image coder included in [12]. The biorthogonal 2D spatial wavelet transform in the JPEG2000 has two main functionalities in our scalabale MVC: Firstly, it helps to reduce the remaining spatial correlation that is still present, particulary in temporal-view-directional lowpass frames and thus enables efficient coding. Secondly, it allows spatial scalability of the multi-view data.

As JPEG2000 is an image codec that produces an SNR-scalable bitstream, we can use it for SNR-scalabilty of multi-view sequences. The associated rate allocation problem is discussed in the following.

The goal of rate allocation in our case is to assign a certain number of bits R_i out of a global bitrate R to each of the $M+1$ subbands in such a way that the resulting distortion D, calculated as MSE between original and reconstructed, is minimal. For orthonormal subband coding, D can be expressed as sum of subband distortions $D_i(R_i)$, where R_i means the bitrate assigned to the subband i. This is because of the energy preservation property of orthonormal transforms. The relation between overall distortion and subband distortion can be extended to biorthogonal transforms by weighting D_i with a weight factor w_i. w_i can be calculated as l_2-norm of the single equivalent reconstruction filter that takes the subband coefficients directly to the reconstructed output [14]. Thus, the problem can be mathematically stated as determination of the optimal rate vector $\mathbf{R}^* = \{R_i^*\}_{i=0}^M$ that minimizes

$$D(\mathbf{R}) = \sum_{i=0}^{M} w_i D_i(R_i) \text{ under the constraint } \sum_{i=0}^{M} R_i \leq R. \tag{2}$$

With Langrangian optimization, this is equal to the minimization of a cost function

$$J(R_i) = D(R_i) + \frac{\lambda}{w_i} R_i \qquad \forall i \in \{0, \dots, M\} \tag{3}$$

where λ is the Lagrange multiplier. Derivation with respect to R_i and setting to zero results in

$$\frac{\partial D_i}{\partial R_i}(R_i) = -\frac{\lambda}{w_i} \qquad \forall i \in \{0, \dots, M\}. \tag{4}$$

Thus, for optimal rate allocation, the points with equal slope on the weighted R-D-curves have to be selected.

In Fig. 3, one can see on the left the R-D testpoints and its approximation by a cubic smoothing spline for one subband after MCTF and DCVF. We use JPEG2000 compression with 20 quality layers for each subband. The figure on the right hand side shows the derivative that can be computed from the spline parameters. In most cases, the curve obtained by spline interpolation gives more robust results (less irregularities) than a "real" R-D-curve obtained from some hundred testpoints would do [13].

III. Experimental Results

A. Coder Description

A block diagram of the proposed coding scheme can be seen in Fig. 4. Both, MCTF and DCVF, use Haar filtering so far, but extension to longer filter kernels in lifting implementation is possible with minor changes. In our experiments we use three decomposition levels for temporal as well as view-directional decomposition.

Motion estimation and disparity estimation use both the same block matching algorithm with block sizes 16x16 and $\frac{1}{2}$-pixel precision. Search ranges can be adapted to the decomposition levels and can differ for horizontal and vertical direction, especially to cope with the geometry of the camera array, when doing disparity estimation. SAD is used to determine the best matching block.

JPEG2000-coding is done with the standard conformant coder from [12]. 20 quality layers are used for all frames

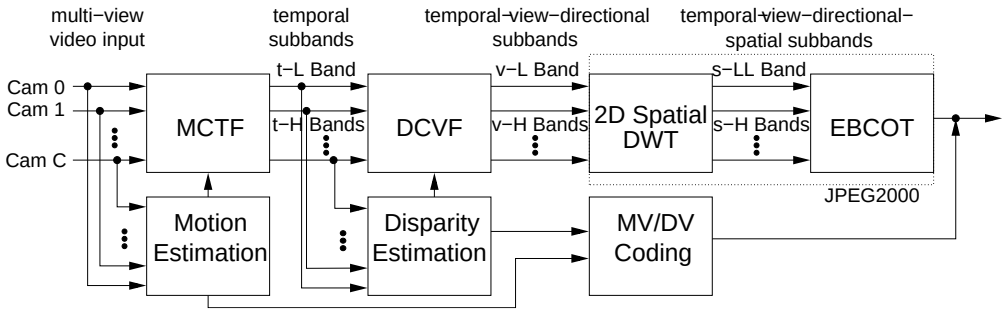


Fig. 4. Block diagram of proposed coding scheme

TABLE I
TRANSFORM CODING GAIN

	G_M	G_{MD}	$\frac{G_{MD}}{G_M}$
race1	54.31	94.85	1.75
race2	40.01	64.85	1.62
golf1	193.11	321.12	1.66

TABLE II
CORRECTED TRANSFORM CODING GAIN AND MV/DV RATES

	$R_M[bpp]$	$R_{MD}[bpp]$	G'_M	G'_{MD}	$\frac{G'_{MD}}{G'_M}$
race1	0.013	0.070	53.33	86.12	1.61
race2	0.032	0.094	38.26	56.95	1.49
golf1	0.009	0.070	190.81	291.54	1.53

of all subbands. All 3 components of the YUV colorspace are compressed at once and so far lossy to lossless coding is performed with 5/3 biorthogonal wavelet filters. Besides of that, the coder is run with its default configuration.

Coding of MVs and DVs is done via exponential-Golomb coding as in H.264.

B. Performance evaluation

To get some primary results, the evaluation of the coder has been performed on the standard multi-view test sequences *race1*, *race2* and *golf1*. All sequences have 8 views, captured by 1D parallel camera arrays, frame rates are 30 fps and spatial resolution 320x240 pixels (QVGA).

In order to get a comparison about how the proposed MCTF-DCVF scheme performs, compared to traditional scalable coders, that only apply an MCTF (e.g. state-of-art 3D simulcast coding), we evaluate the respective coding gains according to [12]. The coding gain denotes the ratio between the distortion that is introduced to a signal when scalar quantization and coding is applied directly on the original source samples (D_{PCM}) or the transform coefficients (D_{TRANS}), respectively, provided that the bitrates are identical.

$$G_{\mathrm{Tranform}} = \frac{D_{\mathrm{PCM}}}{D_{\mathrm{TRANS}}} = \frac{\sigma_X^2}{\prod_{b=0}^{M}(\sigma_{Y_b}^2)^{q_b}}. \qquad (5)$$

Here σ_X^2 denotes the variance of the input signal, that equals the arithmetic mean of subband variances for orthonormal transforms. $\sigma_{Y_b}^2$ is the variance of subband Y_b and q_b denotes the fraction of coefficients in subband b of the total numbers of samples. Thus, the right part in (5) is the ratio between arithmetic and geometric mean of subband samples, that can be interpreted as the energy compaction factor of the transform, which approximates the possible coding gain for scalar quantization of the samples. Table I shows the energy compaction for the three test sequences for MCTF (G_M) and MCTF+DCVF (G_{MD}) transform. The higher values for the latter case demonstrate the superiority of this transform compared to simulcast.

In order to take also into account the motion and disparity information for our coding gain estimation, we multiply formula (5) with a correction factor:

$$G'_{\mathrm{Tranform}} = \frac{D_{\mathrm{PCM}}}{D_{\mathrm{TRANS}}} = \frac{\sigma_X^2 \cdot 2^{-2R_S}}{\prod_{b=0}^{M}(\sigma_{Y_b}^2)^{q_b}}, \qquad (6)$$

where R_S stands for the bitrate that is needed to encode the MVs or MVs+DVs, respectively. As scalar quantization of the samples is assumed in the coding gain formulation, the correction factor is just the additional distortion to D_{TRANS}, that is introduced by the fact that the transform coefficients must be coded with a bitrate that is R_S lower than the original samples. Table II shows the corrected results for the coding gain and the bitrates that are needed to code MVs (R_M) or MVs+DVs (R_{MD}). Obviously the coding gain is lower when taking into account the motion cost, but the new MCTF+DCVF is still superior to the simulcast case.

Fig. 5, 6 and 7 show some preliminary coding results for several scaled representations of the coded multi-view sequence *race2*. As the main focus in the evaluation of the proposed method lies in the feasibility of scalability, coding has been performed with high bit rates up to nearly lossless compression with JPEG2000 in lossless mode. Thus, the highest rate points show the influence of scaling on the quality of the reconstructed sequences without any side effects from coding, i.e. they constitute an upper bound. Of course, the use of JPEG2000 lossless mode has a negative impact on the overall coding efficiency, but this is not the main objective in this paper. It is worth noting that all results have been obtained from one scalable bitstream without transcoding, by

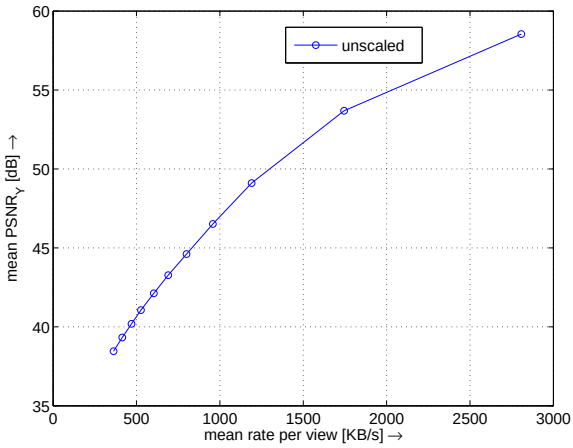


Fig. 5. Mean PSNR$_Y$ for full resolution *race2* sequence

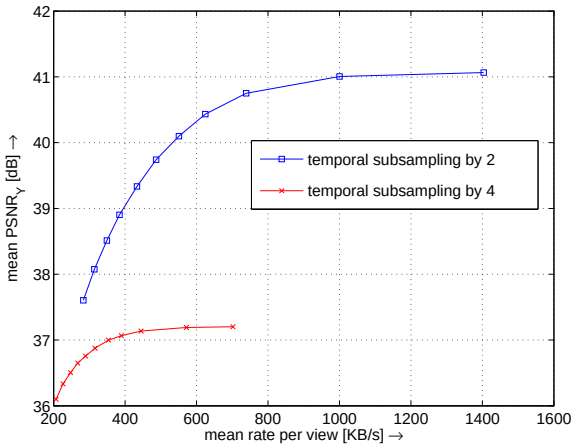


Fig. 6. Mean PSNR$_Y$ for temporally scaled *race2* sequence

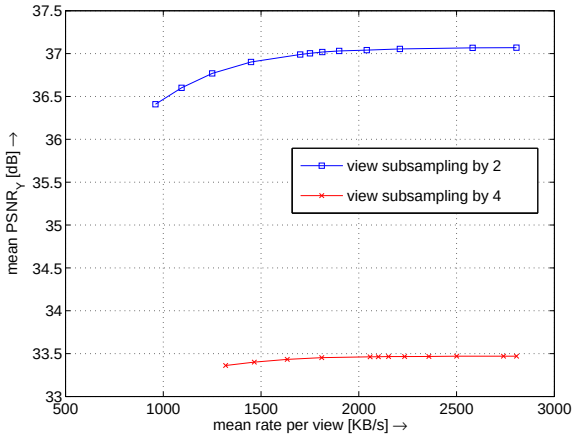


Fig. 7. Mean PSNR$_Y$ for view scaled *race2* sequence

simply discarding unwanted highpass bands or quality layers respectively. Rates are given as average rates per view and PSNR$_Y$ is averaged over all frames and views. It can be seen in Fig. 6 and Fig. 7 that quality degradation caused by scaling is strongly dependent on the decomposition level where data is discarded. E.g. an error introduced by partial reconstruction of the DCVF, also deteriorates performance of inverse MCTF. In general, it can be said that the quality of MC and DC has heavy influence on the quality of scaling.

IV. Conclusion and Future Works

In this paper a wavelet based multi-view video codec has been presented, that provides temporal, view, spatial and SNR scalabilities and uses a new DCVF method. Preliminary simulation results show that the method should provide significant performance gain, compared to simulcast, which has to be evaluated in more detail in the future. Visual quality of scaled sequences is convincing, although motion and disparity compensation are rather simple.

Future work includes the integration of higher order filters in the framework. Different decomposition structures and orders should be investigated. There could also be some improvement by using an entropy coding technique that exploits remaining correlation of the frames inside or even accross the subbands (e.g. 3D ESCOT). More efficient MV and DV estimation and coding is another very important future research topic.

References

[1] S. J. Gortler, R. Grzeszczuk, R. Szeliski, M. F. Cohen, *The Lumigraph*, Proceedings SIGGRAPH 96, pp. 4354, New Orleans, Louisiana, USA, Aug. 49, 1996.

[2] G. J. Sullivan, P. Topiwala, A. Luthra, *The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions*, SPIE Conference on Applications of Digital Image Processing XXVII, Denver, Colorado, USA, Aug. 26, 2004.

[3] U. Fecker, A. Kaup, *Statistical Analysis of Multi-Reference Block Matching for Dynamic Light Field Coding*, Proc. 10th International Fall Workshop Vision, Modeling, and Visualization (VMV 2005), pp. 445-452, Akademische Verlagsgesellschaft Aka GmbH, Erlangen, Germany, Nov. 2005.

[4] K. Mueller, P. Merkle, A. Smolic, T. Wiegand, *Multiview Coding using AVC*, ISO/IEC JTC1/SC29/WG11, MPEG2006/m12945, Bangkok, Thailand, Jan. 2006.

[5] R. Xiong, F. Wu, J. Xu, S. Li, Y-Q. Zhang, *Barbell Lifting Wavelet Transform for Highly Scalable Video Coding*, Proceedings of the Picture Coding Symposium (PCS), San Francisco, USA, Dec. 2004.

[6] J.-R. Ohm, *Three dimensional subband coding with motion compensation*, IEEE Trans. on Image Processing, vol. 3, no 5, pp. 559-571, Sep. 1994.

[7] S. Choi, J. Woods, *Motion Compensated 3-D Subband Coding Of Video*, IEEE Trans. on Image Processing, vol. 8, pp. 155-167, Feb. 1999.

[8] A. Secker, D. Taubman, *Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting*, ICIP 2001, vol. 2, pp. 1029- 1032, Greece 2001.

[9] I. Daubechies and W. Sweldens, *Factoring wavelet transforms into lifting steps*, Journal Fourier Anal. Appl. vol. 4, pp. 247269, 1998.

[10] B. Pesquet-Popescu and V. Bottreau, *Three-dimensional lifting schemes for motion compensated video compression*, Proc. IEEE ICASSP, Salt Lake City, UT, 7-11 Mai 2001

[11] Wenxian Yang, Feng Wu, Yan Lu, Jianfei Cai, King Ngi Ngan, Shipeng Li, *Scalable multiview video coding using wavelet*, IEEE International Symposium on Circuits and Systems (ISCAS), pp. 6078-6081, 2005.

[12] D. Taubmann, M. Marcellin, *JPEG2000–Image Compression Fundamentals, Standards and Practice*, 2nd Edition, Kluwer Academic Publishers, Boston-Dordrecht-London, 2002.

[13] M. Cagnazzo, T. André, M. Antonini, M. Barlaud, *A model-based motion-compensated video coder with JPEG2000 compatibility*, Proceedings of ICIP, Oct. 2004.

[14] B. Usevitch, *Optimal bit allocation for biorthogonal wavelet coding*, Proc. Data Compression Conf, 387-395, Mar. 1996