Barbell Lifting Wavelet Transform for Highly Scalable Video Coding

Ruiqin Xiong[*1], Feng Wu[2], Jizheng Xu[2], Shipeng Li[2], Ya-Qin Zhang[2]*

1) Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
2) Microsoft Research Asia, Beijing, 100080, China

ABSTRACT

This paper proposes a generic lifting technology (Barbell Lifting), where each predicting or updating signal is generated with a Barbell function instead of always a sample from single direction or bi-direction, to incorporate various motion alignment methods into temporal wavelet transform. The proposed lifting technology embraces fractional pixel motion alignment, variable block size motion alignment and overlapped block motion alignment into a framework. It also guarantees the perfect reconstruction in various Barbell functions even with fractional precision. Furthermore, the proposed lifting technology also solves several problems that extensively exist in the current 3D wavelet coding schemes, such as many-to-one mapping for the area with covered scenes, non-referred pixels for the area with uncovered scenes, ambiguities of inversed mapping in the updating stage. Replacing the conventional lifting with the proposed Barbell lifting, the 3D wavelet coding scheme achieves a high coding efficiency.

1. INTRODUCTION

In the application scenario of streaming video over the Internet, video server has to deal with the problems of various device capabilities, diverse network environments and bandwidth fluctuations. A good solution to this problem is to compress the video into a scalable bitstream, which can not only adapt to network variations but also accommodate different devices. Motion aligned 3D wavelet coding schemes provides such solution to keep all the nice scalability features with non-compromised or even better coding efficiency than the state-of-the-art coding schemes.

For simplicity, several early 3D wavelet coding schemes, like [1]~[3], directly decomposed frames temporally along the pixels located at the same position. Due to the widely existing global and/or local motion across frames, the co-located pixels in different frames may be out of alignment so that the temporal wavelet decomposition can not effectively form the high energy compactions, thus reducing the coding efficiency. The motion alignment temporal wavelet decomposition is illustrated in Figure 1, where each column indicates a frame. For example, pixel H0 in F1 is not calculated from the co-located pixels in F0 and F2. Instead, after motion estimation it is decomposed as a high-pass coefficient with the corresponding pixels in F0 and F2 followed the motion trajectory specified by forward and backward motion vectors.

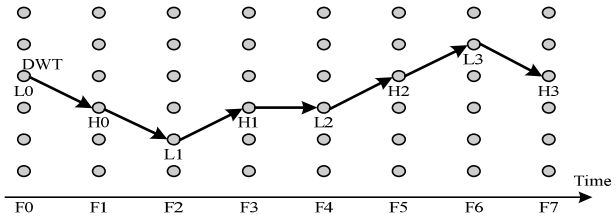


Figure 1: The temporal wavelet decomposition along with motion trajectory.

Many global and local motion models are applied to incorporate motion alignment into the 3D wavelet coding schemes. Taubman et al. [4] pre-distorted video sequence by translating pictures relative to one another before wavelet transform, while Wang et al. [5] used the mosaic technique to warp each video picture into a common coordinate system. Both schemes assume a global motion model, which may be far from the adequacy for many video sequences with local motion. To overcome the limitation, Ohm [6] proposed a block matching technique that is similar to that used in standard video coding schemes while paying special attention to covered / uncovered and connected / unconnected regions. But it fails to achieve perfect reconstruction with motion alignment at the sub-pixel precision. Later, Xu et al. further extended the concept of motion alignment as motion threading for exploiting the long-term correlation across frames along the motion trajectory [7].

Several groups have looked into combining motion alignment with the lifting-based wavelet transform since 2001

[1] This work has been done while the author is with Microsoft Research Asia.

[8]~[13]. One noteworthy work is [8], in which the authors implemented the first sub-pixel (½-pel) resolution motion estimation scheme with perfect reconstruction in the motion-compensated lifting framework. However, the Haar filters were used in [8]. Luo et. al [9] first employed the 5/3 bi-orthogonal wavelet filters with ½-pel resolution motion estimation. In the same year, Secker and Taubman used both the Haar and 5/3 filters again with ½-pel resolution motion estimation [10]. Several follow-up works [11] and [12] also demonstrated the advantage of the 5/3 filters for temporal wavelet transformation.

Further refined motion models with variable block size are also taken into account in the lifting-based temporal wavelet transform. In the MC-EZBC scheme [11][13], motion alignment temporal decomposition is performed by splitting a picture into smaller blocks first and then forming hierarchical variable size block structure. Similar to H.264 [14], Xiong et al also proposed multiple modes with different block sizes for temporal decomposition [15]. Although the motion alignment with variable block size significantly reduces the mean squared error (MSE) of the high-pass frames, such technique also results in many large magnitude coefficients in the subsequent spatial transform because the temporal high-pass pictures contain more edges and blocking artifacts caused by smaller blocks. The technique of overlap-block motion compensation [16] can be borrowed into the lift-based wavelet transform as well. The joint technique of variable block size and overlap-block motion alignment has shown a better performance in [15][17].

However, there are many difficulties in the current temporal wavelet transform. In most of previous works, the 3D wavelet transform is implemented by simply applying classical 1D wavelet transform, convolution based or lifting based, in the motion trajectory, where it is performed as if it were in a 1D signal space. This demands an invertible one-to-one mapping between adjacent frames. Unfortunately, the demand is usually contradictory with the inherent property of motion due to the covered and uncovered scenes within video and the complication of motion.

There are lots of many-to-one mapped and non-referred pixels in the reference frames as shown in Figure 2 (a) and (b). These kinds of relations beyond one-to-one mapping can not be interpreted directly by the classical 1D signal transform even with lifting structure. For example, in the lifting scheme a wavelet filter is usually factored into several predict and update steps. The prediction steps can be carried out as long as the mapping from odd frames to its neighboring even frames is well defined, even if many pixels in an odd frame are mapped to the same pixel in an even frame. However, a reasonable inverse mapping needed for the update steps can not be easily obtained. Most previous works try to define the inverse mapping by using the inverse motion vectors. But for the many-to-one mapped and non-referred pixels in even frames, the inverse mapping for that pixel is ambiguous, as shown in Figure 2 (c).

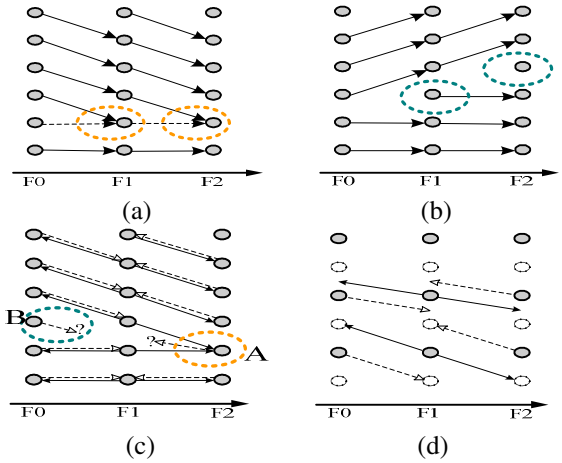


Figure 2: Difficulties in classical temporal wavelet transform.

With fractional pixel motion vectors, the mismatch is even more serious. As shown in Figure 2 (d), pixels in odd frames can be mapping to a "virtual pixel" with a sub-pixel position in even frames. Since the "virtual pixel" can not be updated, is it a right way to assign the inverse motion vector to the nearest integer pixel in the same frame? Furthermore, the action of finding the nearest integer pixel is also ambiguous for half-pixel motion vectors.

The above problems raise the requirement for a new model of 1D transform in multi-dimensional signal space. Instead of using one-to-one mapping, the new model should support many-to-many mapping between pixels of adjacent frames. Therefore, this paper first proposes a general barbell lifting scheme for one dimensional wavelet transform in a multiple dimensional signal space, where the multiple predicting and updating signals are supported through barbell functions. The proposed lifting can embrace all motion alignment techniques mentioned above, such as fractional pixel motion alignment, variable block size motion alignment and overlapped block motion alignment. In particular, it solves those problems that extensively exist in the current 3D wavelet coding schemes. It guarantees the perfect reconstruction in various Barbell functions even at fractional precision.

The rest of this paper is organized as follows. Section 2 describes the proposed barbell lifting. How to solve the existing problems is also discussed in this section. Section 3 discusses the Barbell functions according to various motion alignment techniques. Section 4 describes the proposed 3D wavelet coding with Barbell lifting. Experimental results are given in Section 5. Finally, Section 6 concludes the paper.

2. BARBELL LIFTING

The wavelet decomposition is efficiently implemented by the lifting scheme [18], where every FIR wavelet filter can be factored into lifting stages. The key idea of the lifting scheme contains three steps: the first step splits the data into two subsets, X and Y; the second step uses the subset X to predict the other subset Y and calculates the high-pass wavelet coefficients H as the prediction error; the third step uses the high-pass wavelet coefficients H to update the subset X to ensure preservation of moments in the low-pass L. These steps are briefly referred as Split, Predict and Update step, respectively. In general, only two neighbor elements are used in both the Predict and Update steps. The basic building block in a lifting scheme is a lifting step which takes three input values and generates one through linear combination as follows:

$$y = x_1 + w \times (x_0 + x_2). \quad (1)$$

w is the weighting factor dependent on the wavelet filter.

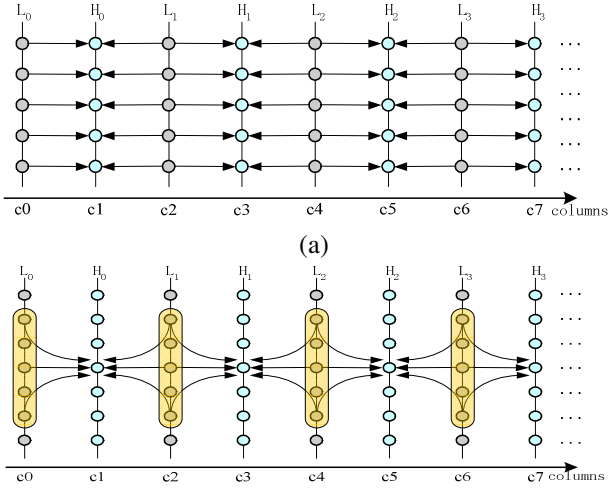


Figure 3: The prediction step of 1D wavelet transform in 2D signal space, (a) conventional lifting; (b) barbell lifting.

To implement an N-dimensional wavelet transform in the N-dimensional signal space, the above one dimensional lifting based wavelet transform has to be applied N times in each of the N directions. Each of the transforms is performed in the same way as if it were in a 1D signal space. For example, Figure 3 (a) shows the Predict step of horizontal wavelet transform in the 2D signal space. Each pixel in odd columns c_{2i+1} is predicted by only one corresponding pixels in each of its two neighboring columns c_{2i} and c_{2i+2}, specified by the transform direction. Other neighboring pixels in columns c_{2i} and c_{2i+2} are not used.

However, when we predict the odd columns c_1, c_3,…, c_{2i+1},…, all the even columns c_0, c_2, …, c_{2i}, … are available. We can expect an even better prediction if we use more available neighboring pixels as sources, as shown in Figure 3 (b). Similarly, to perform a one dimensional wavelet transform in N-dimensional signal space where each c_i corresponds to an N-1 dimensional signal, we can use multiple neighboring pixels in the two N-1 dimensional signal subspaces as prediction source.

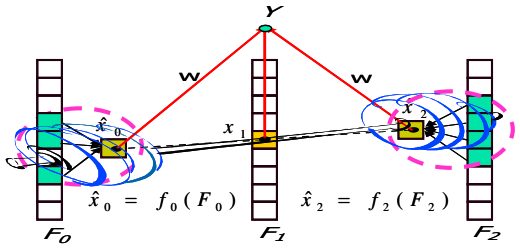
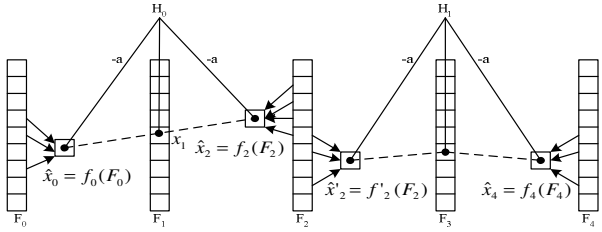
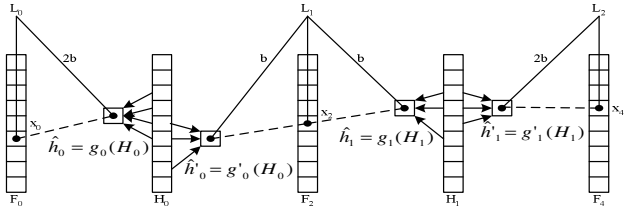


Figure 4: The proposed basic barbell lifting step.

Generally, for a multi-dimensional signal, such as video or image, we propose a novel and generalized lifting scheme for efficient 1D wavelet decomposition in an N-dimensional space. We call it as Barbell lifting scheme as shown in Figure 4, where each of the F_0 and F_2 corresponds to an N-1 dimensional signal subspace. In the proposed Barbell lifting scheme, instead of using a single pixel value, we now use a function of a set pixel values as the input to the lifting step. Functions $f_0()$ and $f_1()$ are called as Barbell functions. They can be any linear or non-linear functions that take any pixels in the F_i as variables. It can also vary from pixel to pixel.



(a) The predict stage



(b) The update stage

Figure 5: The barbell lifting wavelet transform.

The Barbell lifting based forward wavelet transform and inverse wavelet transform are described here. To better explain the idea, assume the input is frames from a video sequence (N=3). Firstly, as shown in Figure 5 (a), the prediction stage takes the original input frames to generate the high-pass frames. The Barbell functions are used to prepare the input values from even frames. Secondly as

shown in Figure 5 (b), the update stage uses the available high-pass frames and even frames to generate the low-pass frames. The Barbell functions are used to prepare the input values from the high-pass frames. For the inverse transform, as long as we know the original Barbell functions used at the update stage, we can first recover the even frames with available high-pass and low-pass frames. If we know the information of the Barbell functions at the prediction stage, we can perfectly reconstruct the odd frames with the available even frames and high-pass frames.

The proposed Barbell lifting can solve the existing problems in the temporal decomposition by coupling the weighting parameters in the updating step with that in the predicting step. Let us denote many-to-many mapping from frame F_i to frame F_j as $\mathcal{M}_{i\to j}$ (many-to-one and one-to-many mapping is dealt as the special case), and $B_{i\to j}(p)\subset F_j$ is the set of pixels in F_j that pixel $p\in F_i$ is mapped to by $\mathcal{M}_{i\to j}$. Define $B^{-1}_{i\to j}(q)=\{p\in F_i \mid q\in B_{i\to j}(p)\}$, $q\in F_j$, which is the set of pixels in F_i that $q\in F_j$ is mapped from by $\mathcal{M}_{i\to j}$. The weighting parameter for pixel pair (p,q) subject to $q\in B_{i\to j}(p)$ is denoted as $w(p,q)$.

The Haar transform with barbell lifting:

$$H_i(p)=F_{2i+1}(p)-\sum_q w(p,q)F_{2i}(q),\ q\in B_{2i+1\to 2i}(p) \quad (2)$$

$$L_i(q)=F_{2i}(q)+\frac{1}{2}\cdot\sum_p w(p,q)H_i(p),\ p\in B^{-1}_{2i+1\to 2i}(q) \quad (3)$$

The 5/3 transform with barbell lifting:

$$H_i(p)=F_{2i+1}(p)-\frac{1}{2}\cdot\left[\sum_{q_1} w(p,q_1)F_{2i}(q_1)+\sum_{q_2} w(p,q_2)F_{2i+2}(q_2)\right],\forall q_1\in B_{2i+1\to 2i}(p),\ \forall q_2\in B_{2i+1\to 2i+2}(p) \quad (4)$$

$$L_i(q)=F_{2i}(q)+\frac{1}{4}\cdot\left[\sum_{p_1} w(p_1,q)H_i(p_1)+\sum_{p_2} w(p_2,q)H_{i-1}(p_2)\right],\forall p_1\in B^{-1}_{2i+1\to 2i}(q),\ \forall p_2\in B^{-1}_{2i-1\to 2i}(q) \quad (5)$$

According to the above barbell lifting process, pixels will be updated by the high-pass coefficients that are predicted with these pixels with the corresponding weighting factor. It guarantees the match in the predicting and updating steps.

3. MOTION ALIGNMENT WITH BARBELL LIFTING

In this section, we will discuss how to decide the Barbell functions for those existing motion alignment techniques used in the 3D wavelet video coding.

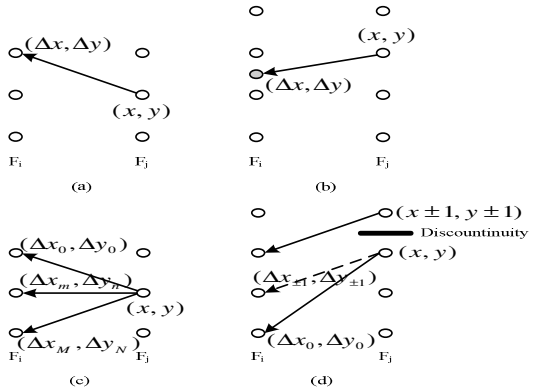


Figure 6: The Barbell functions for temporal decomposition.

Figure 6 gives some examples of Barbell functions. The integer motion alignment case is shown as in Figure 6 (a). The Barbell function is

$$f=F_i(x+\Delta x,y+\Delta y) \quad (6)$$

$(\Delta x,\Delta y)$ is the motion vector of current pixel (x, y). F_i denotes the previous frame. The fractional-pixel motion alignment case is shown as in Figure 6 (b). The Barbell function is

$$f=\sum_m\sum_n w_{m,n}F_i(x+\lfloor\Delta x\rfloor+m,y+\lfloor\Delta y\rfloor+n) \quad (7)$$

$\lfloor\ \rfloor$ denotes the integer part of Δx and Δy. $w_{m,n}$ is the factor of the interpolation filter. In other words, the Barbell function is the fractional pixel value calculated from neighboring integer pixels by an interpolation filter. The multiple-to-one mapping case is shown as in Figure 6 (c). The Barbell function is

$$f=\sum_m\sum_n w_{m,n}F_i(x+\Delta x_m,y+\Delta y_n) \quad (8)$$

$w_{m,n}$ is the weighting factor for each connected pixel.

The Barbell lifting scheme also allows more efficient and flexible temporal decomposition that is not feasible in motion trajectory. Figure 6 (d) shows such a case. Besides the motion vector $(\Delta x,\Delta y)$, the current pixel (x,y) can use the motion vectors of neighboring pixels to get multiple predictions from the previous frame and generate a new prediction. The Barbell function is

$$f=\sum_{m=0,\pm1}\sum_{n=0,\pm1} w_{m,n}F_i(x+\Delta x_m,y+\Delta y_n) \quad (9)$$

$w_{m,n}$ is the weighting factor. Equation (9) describes the case of getting 4 neighboring pixels. It can be extended to more general cases, such as 8 neighboring pixels and even more.

Beside, the Barbell lifting scheme can be also applied with the so-called adaptive block size motion alignment. The Barbell lifting function is used with large block size in the

flat regions of video and small block size in the complex region.

4. THE PROPOSED VIDEO CODING SCHEME

Figure 7 illustrates the block diagram of the proposed 3D wavelet video coding with the Barbell lifting. The raw video is first input to the motion estimation module to estimate the motion among frames. The resulting motion data bases on variable block sizes from 16x16 to 4x4 with quarter pixel precision. The temporal wavelet decomposition uses the proposed Barbell lifting. With the estimated motion data, the high-pass frames are generated by predicting from adjacent frames along with motion trajectory. The overlap motion alignment is also enabled in the predicting step. The updating step uses the coupling weighting parameters as that in the predicting step. After the temporal decomposition, the resulting high-pass and low-pass frames are further decomposed with the 9/7 filter. All sub-bands are coded finally with 3D EBCOT.

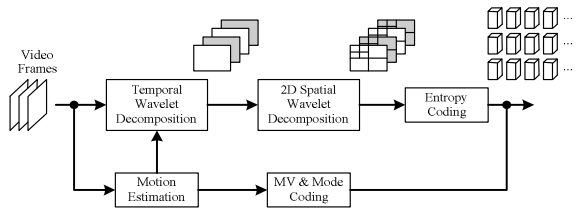


Figure 7: The block diagram of the proposed 3D sub-band video coding using Barbell.

The scheme is submitted to MPEG to response MPEG SVC call for proposals. For the testing scenario with three-layer spatial scalability, the performance of our scheme ranks the first in all submitted scheme.

5. EXPERIMENTAL RESULTS

We have conducted several experiments to show the performance of our scheme in this paper. MPEG standard test sequences are used: *Foreman*, *Mobile*, *Stefan*, *Coastguard* and *Table tennis*. In the tests, the entire sequence is temporally decomposed by a four-level lifting structure into five temporal subbands. Each temporal sub-band is further spatially decomposed by a 3-level Spacl wavelet transform. The resulted wavelet coefficients are entropy coded by 3-D ESCOT. Motion estimation is performed on each level at quarter-pixel accuracy and the motion search range at each level is set as 32, 64, 128 and 128 pixels, respectively. Overlapped block motion alignment with adaptive block size is applied.

We compared our barbell lifting scheme with two benchmark coders: MC-EZBC [13] and H.264 JM6.1e. The result of MC-EZBC is quoted from [19]. For H.264, we try to set the test conditions as follows to obtain its best performance. The GOP is set as a whole sequence with only one I frame; and two B pictures are inserted between each two P pictures; the quantization parameters for these three picture types satisfy $QP_I=QP_P-1=QP_B-2$; motion estimation is performed at quarter-pixel accuracy with a search range of 32; five references are allowed for both the P frames and B frames; and CABAC and R-D optimization are also turned on. From the results shown below, we can see that the proposed scheme outperforms MC-EZBC for all these sequences by 1~2dB. The proposed scheme is also compared with the best results of H.264. For *Foreman* sequences, the loss of our coder is about 0.4 to 1.4 dB. For *Table tennis, Stefan* and *Mobile* sequences, the performance of our coder catches up with that of H.264. Furthermore, for *Coastguard* sequence, our coder even outperforms the best result of H.264 by up to 1.1dB. Considering that the results of H.264 are of single layer bitstreams which are R-D optimized for each bit-rate, our proposed scalable coder is very competitive with H.264.

6. CONCLUSIONS

This paper proposes a general barbell lifting scheme for one dimensional wavelet transform in a multiple dimensional signal space, where the multiple predicting and updating signals are supported through barbell functions. It can embrace all motion alignment techniques mentioned above, such as fractional pixel motion alignment, variable block size motion alignment and overlapped block motion alignment. In particular, it solves those problems that extensively exist in the current 3D wavelet coding schemes. It guarantees the perfect reconstruction in various Barbell functions even at fractional precision.

References

[1]. G. Karlsson and M. Vetterli, "Three dimensional subband coding of video," ICASSP, vol. 2, pp 1100-1103, New York, 1988.

[2]. C. Podilchuk, N. Jayant, and N. Farvardin, "Three-dimensional subband coding of video," IEEE Trans. on Image Processing, vol. 4, no 2, pp. 125-139, 1995.

[3]. Y. Chen and W. Pearlman, "Three-dimensional subband coding of video using the zero-tree method," SPIE on Visual Communications and Image Processing, vol. 2727, pp 1302-1309, 1996.

[4]. D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video", IEEE Trans. Image Processing, vol. 3, no 5, pp. 572-588, 1994.

[5]. Wang, Z. Xiong, P.A. Chou, and S. Mehrotra, "Three-dimensional wavelet coding of video with global motion compensation", DCC, pp 404-413, Snowbird, 1999.

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

[7]. J. Xu, Z. Xiong, S. Li, Y.-Q. Zhang, Memory-constrained 3D wavelet transform for video coding without boundary

effects, IEEE Transactions on Circuits and Systems for Video Technology. vol. 12, no 9, pp 812–818, 2002.

[8]. B. Pesquet-Popescu, and V. Bottreau, “Three-dimensional lifting schemes for motion compensated video compression,” ICASSP, vol. 3, pp 1793–1796, Salt Lake City, 2001.

[9]. L. Luo, J. Li, S. Li, Z. Zhuang, and Y.-Q. Zhang, “Motion compensated lifting wavelet and its application in video coding,” ICME, pp 365-368, Tokyo, 2001.

[10]. Secker, and 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.

[11]. P. Chen and J. Woods, “Bidirectional MC-EZBC with lifting implementation,” IEEE Transactions on Circuits and Systems for Video Technology, to appear.

[12]. M. Flierl and B. Girod, “Video coding with motion-compensated lifted wavelet transforms,” submitted to Signal processing: Image Communication, 2003.

[13]. P. Chen and J. W. Woods, “Improved MC-EZBC with quarter-pixel motion vectors”, MPEG document, ISO/IEC JTC1/SC29/WG11, MPEG2002/M8366, Fairfax, VA, May 2002.

[14]. T. Wiegand, G. Sullivan, G. Bjntegaard and A. Luthra, “Overview of the H.264/AVC video coding standard,” IEEE trans. on CSVT, vol. 13, no. 7, pp 560-576, 2003.

[15]. R. Xiong, F. Wu, S. Li, Z. Xiong and Y.-Q. Zhang, “Exploiting temporal correlation with adaptive block-size motion alignment for 3D wavelet coding,” SPIE, vol. 5308, pp 144-155, 2004.

[16]. M.T. Orchard and G.J. Sullivan, “Overlapped block motion compensation: an estimation-theoretic approach,” IEEE trans. Image Processing, vol. 3, no 5, pp 693-699, 1994.

[17]. Y. Wu and J. W. Woods, “Recent improvements in the MC-EZBC Video Coder,” MPEG document, ISO/IEC JTC1/SC29/WG11, MPEG2003/M10396, Hawaii, 2003.

[18]. Daubechies and W. Sweldens, “Factoring wavelet transforms into lifting steps,” Journal Fourier Anal. Appl. vol. 4, pp 247–269, 1998.

[19]. P.Chen, and J.W.Woods, “Exploration Experimental Results and Software”, JVT proposal, ISO/IEC JTC/SC29/WG11, MPEG2002/M8524, Klagenfurt, AT, July 2002.