3D BASED VIDEO CODING IN THE OVERCOMPLETE DISCRETE WAVELET TRANSFORM DOMAIN WITH REDUCED DELAY REQUIREMENTS

Vidhya Seran, Lisimachos P. Kondi

332 Bonner Hall, Dept. of Electrical Engineering
State University of New York at Buffalo, Buffalo, NY 14260
Tel: (716) 645-2422 ext. 2133 Fax: (716) 645-3656
Email: { vseran,lkondi }@eng.buffalo.edu

ABSTRACT

In this work, we propose a new temporal filter set to minimize delay in 3D wavelet based video coding, which gives a performance at par with existing longer filters. Our filter set will not have any boundary effects at the group of frames (GOF). The length of the GOF can vary from five to any number of frames depending on delay requirements. We also propose a novel technique of assigning priorities to temporal subbands at different levels to control distortion fluctuation inside a GOF. Experimental results are presented and conclusions are drawn.

1. INTRODUCTION

All current video compression standards are based on the Motion-Compensated Discrete Cosine Transform (MC-DCT) paradigm and its variations. This paradigm has been in use for over two decades and is widely used in a wide range of applications. However, wavelet-based compression is known to outperform DCT-based compression for image coding and the JPEG-2000 image compression standard is wavelet-based. In order to have efficient video compression, the temporal redundancy in the video data has to be properly exploited. Initial approaches to applying motion compensation to the Discrete Wavelet Transform (DWT) were not very successful. If motion compensation is performed in the spatial domain, as in MC-DCT based codecs, and the prediction error is encoded using DWT instead of DCT, compression efficiency will not be good since the DWT is not well suited to the statistics of the prediction error. Also, band-to-band motion compensation in the DWT domain is not efficient because the DWT is not shift-invariant and the wavelet coefficients of the current frame cannot be accurately predicted from the coefficients of the previous frame.There are two main theoretical developments that promise efficient wavelet-based video codecs: Temporal filtering using lifting , and motion compensation in the Overcomplete Discrete Wavelet Transform (ODWT) domain.

3D wavelet-based video coding schemes employ a three dimensional wavelet transform. Thus, temporal redundancy in the video source is exploited using temporal filtering. 3D schemes offer drift-free scalability. The tradeoff is an increase in delay requirements, since, in contrast to 2D methods, frames cannot be encoded one by one but processing is done in groups of frames. Thus, a number of frames need to be available to the encoder before coding can begin. Furthermore, the whole group of frames needs to be received by the receiver before decoding can start. Thus, 3D video coding schemes offer better performance but also relax the causality of the system.

The three-dimensional wavelet decomposition can be performed in two ways: two-dimensional spatial filtering followed by temporal filtering (2D+t) [3, 1, 2] or, temporal filtering followed by two-dimensional spatial filtering (t+2D) [4, 5, 6]. We propose to perform the two-dimensional spatial filtering first and then perform motion-compensated temporal filtering using lifting in the Overcomplete Discrete Wavelet Transform domain.

The rest of the paper is organized as follows: In Section 2, we explain the motion-compensated temporal filtering (MCTF) using lifting and related problems. In Section 3, we discuss our proposed methods: Section 3.1 explains our new filter set to minimize delay and Section 3.2 addresses bit allocation for temporal subbands. Finally, in Section 4, we present the simulation results.

2. MOTION-COMPENSATED TEMPORAL WAVELET TRANSFORM USING LIFTING

Lifting allows the incorporation of motion compensation in temporal wavelet transforms while still guaranteeing perfect reconstruction. Any Finite Impulse Response (FIR) filter can be implemented using lifting. Let us consider as an example the Haar wavelet transform:

$$\begin{aligned} h_k(x,y) &= f_{2k+1}(x,y) - f_{2k}(x,y) \\ l_k(x,y) &= \frac{1}{2}[f_{2k}(x,y) + f_{2k+1}(x,y)], \end{aligned} \tag{1}$$

where $f_k(x,y)$ denotes frame k and $h_k(x,y)$ and $l_k(x,y)$ represent the high-pass and low-pass subband frames. Using lifting, the Haar filter can be implemented as:

$$\begin{aligned} h_k(x,y) &= f_{2k+1}(x,y) - f_{2k}(x,y) \\ l_k(x,y) &= f_{2k}(x,y) + \frac{1}{2}h_k(x,y). \end{aligned} \tag{2}$$

It is possible to modify the above equations in order to incorporate motion compensation. Let $W_{i\to j}(f_i)$ denote the motion-compensated mapping of frame f_i into frame f_j. Thus, the operator $W_{i\to j}(.)$ gives a per pixel mapping between two frames. No particular motion model is assumed. Thus, the above equations are modified as

$$\begin{aligned} h_k(x,y) &= f_{2k+1}(x,y) - W_{2k\to 2k+1}(f_{2k})(x,y) \\ l_k(x,y) &= f_{2k}(x,y) + \frac{1}{2}W_{2k+1\to 2k}(h_k)(x,y). \end{aligned} \tag{3}$$

0-7803-9134-9/05/$20.00 ©2005 IEEE

These equations correspond to taking the Haar transform along the motion trajectories [7].

As mentioned previously, instead of the Haar transform, any two-channel FIR subband transform can be implemented using motion-compensated lifting. For the case of the biorthogonal 5/3 wavelet transform, the analysis equations are

$$\begin{aligned} h_k(x,y) &= f_{2k+1}(x,y) - \frac{1}{2}(W_{2k\to 2k+1}(f_{2k})(x,y) \\ &\quad + W_{2k+2\to 2k+1}(f_{2k+2})(x,y)) \\ l_k(x,y) &= f_{2k}(x,y) + \frac{1}{4}(W_{2k-1\to 2k}(h_{k-1})(x,y) \\ &\quad + W_{2k+1\to 2k}(h_k)(x,y)). \end{aligned} \tag{4}$$

In the lifting operation, the prediction residues (temporal high-pass subbands) are used to update the reference frame to obtain a temporal low subband. We will refer this as update step in the following discussions.

If the motion is modeled poorly, the update step will exhibit ghosting artifacts to the lowpass temporal subbands. The update step for longer filter depends on more number of future frames. Also, the grouping of video frames of finite size becomes difficult for longer filters. If the entire sequence is treated as a single Group of Frames (GOF), then this approach is reasonable for algorithm simulation but not for real world applications.

If a video sequence is divided into a number of GOFs that are processed independently, without using frames from other GOFs, when 5/3 lifting is used, high distortion will be introduced at the GOF boundaries. The distortion will be in the range of 4-6 dB (PSNR) at the GOF boundaries irrespective of the motion content or model used [6, 1, 8]. To reduce this variation at the boundaries, we need to use frames from past and future GOFs. Thus, it is clear that the introduced delay (in frames) is greater than the number of frames in the GOF. The encoding and decoding delay will be very high as the encoder has to wait for future GOFs. In [6], the distortion at the boundaries is reduced to some extent by using a sliding window approach. This method with GOF=8 frames, needs the current GOF plus 14 frames, for a three level 5/3 temporal decomposition. For an additional level of temporal decomposition, the delay will be almost doubled. We should note that, even when the delay is high, the low pass temporal subbands are not free from ghosting artifacts.

In temporal filtering, we also notice distortion variation within a GOF, which is not directly related to the motion in the sequence as in the case of hybrid video coding schemes. The distortion fluctuation is more pronounced in longer filters and is undesirable at low bitrates [6, 1].

By skipping entirely the update step for 5/3 filter [7, 9], the analysis equations can be modified as,

$$\begin{aligned} h_k(x,y) &= f_{2k+1}(x,y) - \frac{1}{2}(W_{2k\to 2k+1}(f_{2k})(x,y) \\ &\quad + W_{2k+2\to 2k+1}(f_{2k+2})(x,y)) \\ l_k(x,y) &= f_{2k}(x,y). \end{aligned} \tag{5}$$

We refer to this filter set as 1/3 transform.

Filters without update step, will allow us to group frames of finite size and with less ghosting artifacts in the low pass temporal subbands. Hence, by avoiding the update step, we get high quality temporal scalability. But at full frame rate resolution, the 1/3 filter suffers in compression efficiency compared to the 5/3 filter.

So far, an overview of motion compensated temporal filtering and the problems were discussed.

3. PROPOSED 3D-CODER

Our proposed 3D-coder addresses the following :

- Design of Temporal Filtering Schemes to Minimize Delay Requirements
- Bit Allocation Between Temporal Subbands.

3.1. Design of Temporal Filtering Schemes to Minimize Delay Requirements

In 3D coding schemes, high level of compression is achieved by applying temporal filter for a group of frames. The number of frames in a buffer will increase with the length of the filter and the number of temporal decomposition levels. This introduces a delay both at the encoder and decoder.

We propose a new filter set that minimizes delay and performs at par with longer filters. In this filter design, lowpass temporal frames are created at the beginning and at the end of a group of frame at the first level temporal decompostion. Unlike Haar or longer filters like 5/3, which require GOFs to be in some power of 2, the proposed filter does not have any constraints on GOF length.The proposed filter set is perfectly invertible. This can be applied for both t+2D and 2D+t schemes.

We first describe the proposed filter design without including any update step. Thus, the lowpass temporal frames are unfiltered original video frames. Let N be the length of the GOF and L be the maximum number of temporal decomposition levels. Let l ($1 \leq l \leq L$) be the l^{th} level temporal decomposition. At any level l, if the number of low pass temporal subband within a GOF is greater than 2, bi-directional motion estimation is used to evaluate highpass temporal subbands. If the number of lowpass temporal subband is equal to 2, we can restrict ourselves to forward estimation within the GOF or perform bi-directional estimation by considering a lowpass temporal subband from the next GOF. The lowpass temporal subband also happens to be the first frame of the next GOF. Hence we need only the GOF or GOF plus one future frame. At l^{th} level decomposition, if there is only one low pass temporal subband and $l < L$, then we apply forward estimation with adjacent GOF. The proposed filter set is pictorially represented for GOF=5 in Figure 1.

At level $l = 1$, the lowpass temporal subbands are placed at the beginning and the end of the GOF. If N is even, we will have two sucessive lowpass subbands at tht end of the GOF (refer to figure 2. In figure 1, at level $l = 2$ we have two lowpass frames. So, to estimate the LLH, we can either use the first frame from the next GOF (LL) to do bi-directional motion estimation or just use the forward estimation. At $l = 3$, we have only one LLL frame and hence forward estimation is carried out using the LLL_3 from the next GOF. If GOF=8, we get two lowpass frames at $l = 3$. We can make the GOF totally independent without using any future GOFs or we can just take one future frame (refer to figure 2). One should note that four levels of temporal decomposition in the proposed filter set actually correspond to three levels in Haar or longer filters, in the sense that the same number of lowpass frames exist in both the cases.

Also, Figure 2 is used to explain our proposed scheme for GOF=8. At level 3, in figure 2, as we discussed earlier, we have

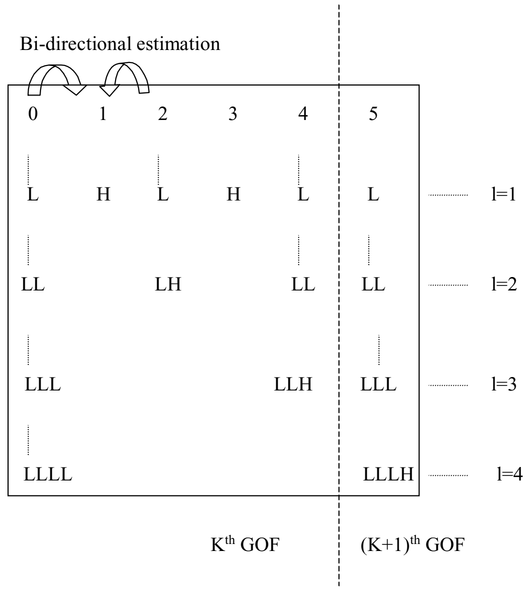

Fig. 1. Proposed Filter for GOF=5.

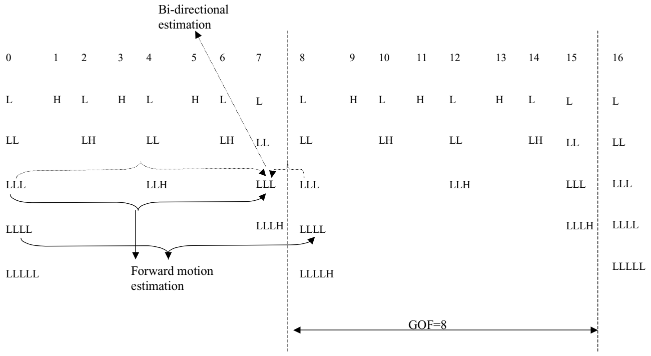

Fig. 2. Proposed Filter with GOF=8

two options: forward estimation (solid line) or bi-directional estimation (dashed line).

To summarize, when the GOF length is less than 8, we need only one future frame to get a compression efficiency equal to that of 5/3 filters. If the GOF length is greater than or equal to 8, the GOF can be coded without the knowledge of past and future GOFs. This gives us flexibility in choosing the GOF length, based on delay constraints and compression requirements. For the proposed case, when GOF=5 frames and $L = 4$, then the initial coding delay will be GOF+1 frames and it yields a performance comparable to that of 1/3 filters, which require nine frames for a three-level temporal decomposition. The longer the GOF length better will be the compression efficiency. The total number of motion vectors for the proposed filter set will never exceed 1/3 filter case.

The lifting update step can be included in the proposed filter to increase the overall compression efficiency without increasing the delay. In our proposed filter set, inclusion of an update step will not change the delay. It will remain the same as in the case without update. Our update procedure varies at different temporal decomposition level and the design is such that, it will not use future highpass temporal subbands from another GOF.

For our proposed case, if we increase our GOF length, we will obtain better compression without increasing delay. The boundary effects will not be seen in the proposed filter set.

3.2. Bit Allocation Between Temporal Subbands

All current wavelet-based video codecs that employ temporal filtering exhibit a significant fluctuation in the PSNRs of the frame within a GOF. This is true for both t+2D and 2D+t schemes. The distortion fluctuation inside a GOF can be in the order of 1-4 dB [1, 6]. It is well known that the average PSNR for the whole video sequence alone is not an adequate indicator of subjective video quality and the PSNR fluctuation should be taken into account. Hence, the PSNR variation inside a GOF should be controlled.

We propose a novel technique of assigning priorities to temporal subbands at different levels to control distortion fluctuation inside a GOF. Some temporal subbands will have larger impact on the frames inside a GOF and take precedence over others. For example, in a three level temporal decomposition for a Haar filter for GOF of eight, the third level filtered temporal low and high bands will have an effect on the entire GOF, while the first level filtered temporal high frames will affect only one frame. For this example, the third level lowpass temporal subbands should be treated like an intra frame in hybrid video coders. Also, the higher the temporal level, the higher the energy in the high pass temporal subbands. This is because the distance between the frames get doubled at each higher temporal level. Hence, different temporal subbands will have different energy content and should be treated differently. This proposed technique will also minimize quantization error propagation from one level to another. In addition, at different temporal resolutions we get a high quality output.

We propose a bit allocation procedure to determine the bits to be encoded for each temporal subband. Let R_{GOF} be the rate allocated for a GOF, R_{Low_L} be the rate for low pass temporal subband and R_{H_l} be the rate for high pass temporal subband at level $0 \leq l \leq L$. Then ,

$$R_{Low_L} > R_{H_L} > R_{H_{L-1}} > \dots R_{H_l} > \dots R_{H_1} \qquad (6)$$

Our experimental results have shown, that the PSNR fluctuations within a GOF are greatly reduced for any type of filter used.

4. EXPERIMENTAL RESULTS

We have considered four standard test sequences, two in SIF (352×240) resolution: “Football” and “Flower Garden” and two in QCIF (176×144) resolution: “Foreman” and “Susie”. A Daubechies $(9, 7)$ filter with a three level decomposition is used to compute the wavelet coefficients. The motion estimation is performed in the overcomplete wavelet domain using the block matching technique for integer pixel accuracy. A 16×16 wavelet block is matched in a search window of $[-16, 16]$.

In the first set of experimental results, we gauge the performance of the proposed temporal filters that offer reduced delay

Table 1. Average PSNR values of Y component for 0.5bpp.

Sequence	Haar w/o update	Haar with update	1/3 Filter	Proposed Filter (GOF=5)	Proposed Filter (GOF=8)
Football	27.63 dB	28.75 dB	29.42 dB	29.24 dB	29.96 dB
Foreman	35.26 dB	36.02 dB	37.41 dB	37.11 dB	37.66 dB
Susie	41.18 dB	42.81 dB	43.07 dB	42.98 dB	43.19 dB

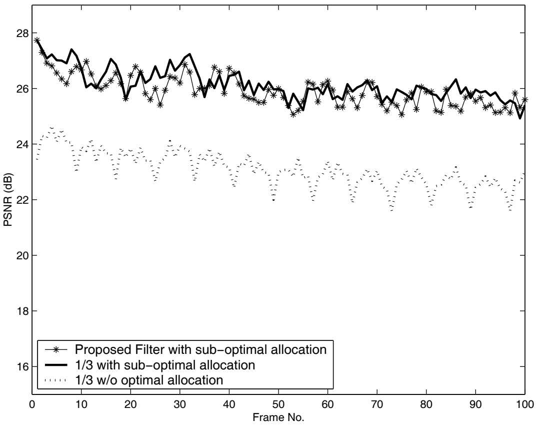

Fig. 3. Flower garden at 1Mbps .

requirements. The temporal subbands are compressed using 3D-SPIHT coder [11]. The proposed temporal filter (version without update) with a GOF of five and eight frames are compared with 1/3 and Haar filters with a GOF of eight frames. Table 1 gives the average PSNR values of the Y component for an encoded bit rate of 0.5 bpp. It can be seen from the Table 1 that the proposed filter outperforms the Haar filters and is competitive with the 1/3 filter, while having lower delay requirements. The proposed filter with GOF=8, performs better than 1/3 filter.

In the second set of experimental results, we show the importance of bit allocation across temporal subbands. A suboptimal bit allocation was tried that satisfied Eq. (6). The rates in bits per pixel (bpp) used for selecting the bit allocation are

$$R_{Low_3} = 1.0 > R_{H_3} = 0.63 > R_{H_2} = 0.45 > R_{H_1} = 0.36 \quad (7)$$

The 2D-SPIHT image coder [10] was used to encode each temporal subband independently so that we could easily select the number of bits to be used for each temporal subband. The “Flower Garden” sequence was encoded at 1 Mbps using 1/3 filter without bit allocation and the 1/3 and proposed filters with bit allocation. The PSNR of each frame is plotted in Figure 3. It can be seen that, with the bit allocation scheme the PSNR variation is greatly reduced and the average PSNR is also increased.

5. CONCLUSION

We have proposed a novel temporal filter set with motion compensation for 3D wavelet-based video coding along with a scheme for bit allocation across temporal subband. The filter set described offers flexible features for compression efficiency and delay requirements. Our experimental results show, the effectiveness of the proposed scheme.

6. REFERENCES

[1] Y. Wang, S. Cui, and J. E. Fowler, “3D video coding using redundant-wavelet multihypothesis and motion-compensated temporal filtering,” in *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain, 2003, vol. 2, pp. 755–758.

[2] Xin Li, “Scalable video compression via overcomplete motion compensated wavelet coding,” *Signal Processing: Image Communication (special issue on ”Subband/Wavelet Interframe Video Coding”)*, vol. 19, pp. 637–651, August 2004.

[3] Y. Andreopoulos, A. Munteanu, J. Barbarien, M. van der Schaar, J. Cornelis, and P. Schelkens, “In-band motion compensated temporal filtering,” *Signal Processing: Image Communication (special issue on ”Subband/Wavelet Interframe Video Coding”)*, vol. 19, pp. 653–673, August 2004.

[4] D. Taubman and A. Zakhor, “Multirate 3-D subband coding of video,” *IEEE Transactions on Image Processing*, vol. 3, pp. 572–588, Sept. 1994.

[5] S. T. Hsiang and J. W. Woods, “Embedded video coding using motion compensated 3-D subband/wavelet filter bank,” in *Proceedings of the Packet Video Workshop*, Sardinia, Italy, May 2000.

[6] A. Golwelkar and J. Woods, “Scalable video compression using longer motion compensated temporal filters,” in *Proc. SPIE VCIP*, 2003, vol. 5150, pp. 1406–1416.

[7] A. Secker and D. Taubman, “Lifting based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression,” *IEEE Transactions on Image Processing*, vol. 12, pp. 1530–1542, Dec 2003.

[8] A.Golwelkar, *Motion compensated temporal filtering and motion vector coding using longer filters*, Ph.D. thesis, Rensselaer Polytechnic Institute, 2004.

[9] M. Van der Schaar and D. Turaga, “Unconstrained motion compensated temporal filtering(UMCTF) framework for wavelet video coding,” in *Proc.of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 2003.

[10] A. Said and W. Pearlman, “A new, fast, and efficient image codec based on set partitioning in hierarchical trees,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 243–250, June 1996.

[11] B. J. Kim, Z. Xiong, and W. A. Pearlman, “Low bit-rate scalable video coding with 3D set partitioning in hierarchical trees,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 1374–1387, Dec 2000.