MOTION-COMPENSATED TEMPORAL FILTERING BASED ON THE DCT

Randa Atta

Dept. of Electrical Engineering
Suez Canal University
Port Said, Egypt

Mohammad Ghanbari

Audio & Video Networking Research Lab
Dept. of Computing and Electronic Systems
University of Essex, Colchester, UK

ABSTRACT

Motion-compensated temporal filtering (MCTF) based lifting implementations of various discrete wavelet transforms have recently gained a lot of interest due to their good performance in energy compaction and their ability to provide various scalability features. Although all the existing MCTF schemes are based on the wavelet transform, in this paper, we propose a temporal filter framework based on the discrete cosine transform (DCT) which is an extension of our motion compensated DCT temporal filter (MCDCT-TF) [1]. In the current work, in addition to the two-band and three-band temporal decomposition structures employed in the MCDCT-TF technique, a longer tap filter (5/3 DCT) is utilized to improve the compression gain further. Simulation results show that a three-dimensional hybrid 3D subband/DCT codec with longer tap DCT filters yields a significant improvement over our earlier 3/2 MCDCT-TF, Haar, and 5/3 wavelet filters.

Index Terms— Motion compensation, Lifting, temporal discrete wavelet transform, DCT, scalable video coding.

1. INTRODUCTION

Three-dimensional subband video coding is becoming increasingly popular, as it provides high compression performance comparable to the state-of-the-art H.264/AVC codec [2]. It also provides a wider range of scalability features and solutions for network congestion. The 3D subband schemes exploit the temporal redundancy by applying MCTF over the frames of a video sequence. Recently, most of the proposals for implementing MCTF techniques have been based on the wavelet transforms. The subband/wavelet temporal filters are implemented with lifting scheme as it allows introducing non-linear computations, such as motion compensation, and guarantees invertibility for any arbitrary sub-pixel accuracy. A two-channel decomposition can be achieved with a sequence of two successive steps: prediction and update steps that form a lifting structure. In the prediction and update steps the high and low temporal subbands are obtained respectively.

Of the various motion-compensated 3D SBC schemes based on lifting framework reported in the literature [3-6], some use the 2-tap Haar filter (as a short temporal filter) for motion compensated temporal filtering while others use longer length filters like the 5/3 filter to take a better advantage of the temporal redundancy among the frames. In [3] and [6], the longer temporal filters such as 5/3 filters were used with either bi-directional or uni-directional motion estimation to achieve a 3-D subband/wavelet video coding. Their results show that longer filters have higher coding gains and significant PSNR improvement at higher bit rates over the 2-tap Haar filters. The MCTF coding framework not only has been employed in the 3D-wavelet-based video coding but also into the state-of-the-art scalable video coding scheme, the scalable extension of H.264/AVC [7][8]. It has been shown that use of MCTF in the H.264/AVC codec significantly improves its compression and scalability efficiency. However, the MCTF is not a normative part of the SVC since the closed-loop hierarchical B-Frames has shown to be more efficient than the open loop MCTF.

The first step in realization of MCTF with the DCT transform instead of the wavelet transform was introduced in [1]. This temporal filter framework MCDCT-TF not only provides a band partitioning structure that allows generation of low and high subbands but also a two-band structure, required in the lifting process. In this paper, we extend the concept of MCTF in the DCT domain with longer tap filters in order to take even more advantage of temporal redundancy. They are realized in a pyramidal structure with various levels, depending on the filter length and its odd/even tap nature.

2. DCT TEMPORAL ANALYSIS AND SYNTHESIS

An important issue in temporal multiresolution analysis is the choice of the temporal filter length. Longer tap filters take a better advantage of the temporal correlation among the successive frames. In this section, we show how lifting of 5 frames in the MCDCT-TF temporal decomposition structure can be realized, though generalization to longer DCT length is a straightforward extension. It should be noted that the main constraint on the length of set of DCT coefficients, in addition to the encoding delay, is the realization of connected versus unconnected pixels. Hence we first describe the MCDCT-TF temporal analysis, introduced in [1] and then extend the analysis to a larger DCT dimension.

2.1. MCDCT-TF framework

In [1], realization of the MCDCT-TF framework was based on the combinations of three-band and two-band temporal decomposition structures. We call the DCT temporal filter employed in each DCT structure *N*/*L*, where *N* is the number of input frames to be filtered and *L* is the number of low frequency filtered frames, to be retained for further decomposition.

For a three-band structure, three frames are two-band filtered, say the two lower bands into one group and the other band in the next group. We call this as 3/2 DCT filter. The 3/2 DCT filter is applied on each set of three frames denoted by x_{3t}, x_{3t+1} and x_{3t+2}. In this case, there are two motion vectors: forward and backward, denoted by v^f and v^b respectively. Then a one-dimensional (1D) forward 3×1 DCT transform is applied to the aligned three pixels of x_{3t+1} and the two displaced reference pixels x_{3t} and x_{3t+2} as follows:

$$DCT\text{-}3\times 1\ x_{3t}(m+\bar{v}_m^f, n+\bar{v}_n^f), x_{3t+1}(m,n), x_{3t+2}(m+\bar{v}_m^b, n+\bar{v}_n^b) \quad (1)$$

which results in 3 DCT coefficients. We classify them into two lower frequency coefficients L_{2t} and L_{2t+1} and a high frequency coefficient H_t. Now through an inverse transform of 1D 2×1 IDCT of the two lower bands L_{2t} and L_{2t+1}, two lower frequency pixels of l_{2t} and l_{2t+1} are obtained to form the two low-pass subbands. Therefore, the equations that compute the high and two low-pass subbands for connected pixels are given by:

$$\begin{aligned} H_t(\mathrm{m,n}) &= x_{3t+1}(m,n) - \frac{1}{2}\left(\tilde{x}_{3t}(m+v_m^f, n+v_n^f) + \tilde{x}_{3t+2}(m+v_m^b, n+v_n^b)\right) \\ l_{2t}(m+\bar{v}_m^f, n+\bar{v}_n^f) &= x_{3t}(m+\bar{v}_m^f, n+\bar{v}_n^f) \\ l_{2t+1}(m++\bar{v}_m^b, n+\bar{v}_n^b) &= x_{3t+2}(m+\bar{v}_m^b, n+\bar{v}_n^b) \end{aligned} \quad (2)$$

more details on handling the connected and unconnected pixels can be found in [1]. The synthesis procedure is the reverse process of the temporal analysis.

In the two-band structure, successive pairs of frames are temporally filtered using a 2/1 DCT filter to create low-pass (*l*) and high-pass (*H*) frames. This filter is similar to a Haar filter. With the 2/1 DCT filter, first, the 1D 2×1 DCT is applied on every two consecutive frames (in fact one is the predicted frame and the other is motion compensated with respect to the predicted frame). It is then followed by 1D 1×1 IDCT (just one scaled pixel) of the DC coefficient, scaled down by a factor of $\sqrt{2}$ to avoid the dynamic range expansion of the generated low subband frames. The high band becomes the high frequency band of the decomposed frames and the lower band is retained for further decomposition. This filtering operation for the connected pixels can be written as:

$$\begin{aligned} l(m,n) &= (B(m,n) + A(m+v_m, n+v_n))/2 \\ H(m,n) &= (B(m,n) - A(m+v_m, n+v_n))/\sqrt{2} \end{aligned} \quad (3)$$

Unconnected pixels, in frame *A*, are copied into the low subband frames. This is done to avoid the appearance of "holes" in the filtered low band frames. As can be seen from the analysis and synthesis, the 3/2 DCT scheme is invertible at subpixel accuracy but the 2/1 DCT filter is not invertible.

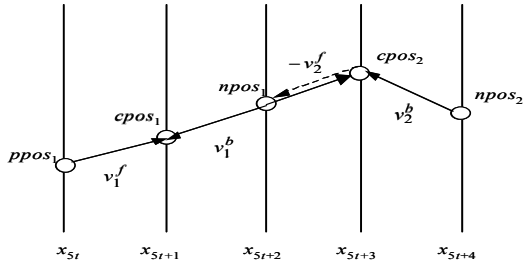

Fig. 1. The positions corresponding to the motion vectors in successive frames for the 5/3 DCT filter.

2.2. Longer tap MCDCT-TF framework

Successive application of forward *N*×1 and inverse *L*×1 transforms of the *L*<*N* retained lower frequency bands leads to the realization of longer tap filters for MCTF. However, the relation between the lengths *N* and *L* at the first stage will put a constraint on the lengths of the forward and inverse transforms at the latter stages. Hence for efficient results, we may have to use variable size forward and inverse transforms at various stages of temporal decomposition. For example, for a 5-tap filter, a group of 5 frames is decomposed into four temporal bands by 3-stage recursive decomposition. The 5/3 DCT temporal filter is first applied to these 5 frames resulting in three low and two high subband frames. After one stage decomposition, a new GOF is made of the three low subband frames and the remaining two high subband frames are the first set of the generated high band frames. This new GOF of 3 lower frequency frames is recursively decomposed into two frames at stage #2 by using the 3/2 DCT filter. The 2/1 DCT is then used to obtain one low and one high subband frames at stage #3.

For the implementation of the 5/3 DCT temporal filter, two bi-directional motion estimations are performed to each sub-GOF of 5 frames. The first bi-directional motion estimation is applied between frame x_{5t+1} and its reference frames x_{5t}, and x_{5t+2}. The obtained forward and backward motion vectors are respectively called v_1^f and v_1^b. The second bi-directional motion estimation is between frame x_{5t+3} and its reference frames x_{5t+2} and x_{5t+4} with the forward and backward motion vectors of v_2^f and v_2^b respectively. Positions corresponding to these motion vectors in successive frames are illustrated in Fig. 1. The 5/3 DCT temporal filter is then applied to the five connected pixels in a sub-GOF of 5 frames. To find these five pixels, we have to determine their positions. First consider a pixel x_{5t+1} located at position $cpos_1$ and the best matched pixels in the previous and next frames (x_{5t} and x_{5t+2}) are located at the nearest pixel positions $ppos_1$ and $npos_1$, respectively. Positions $ppos_1$ and $npos_1$ can be

determined in terms of the current position $cpos_1$ and the two motion vectors v_1^f and v_1^b. The problem is then to find position $cpos_2$, which is associated to $npos_1$. We infer the backward motion vector (shown by the dashed lines in Fig. 1) for the pixel located at position $cpos_2$ in frame x_{5t+3} from the forward motion vector, as ($-v_2^f$). Position $cpos_2$ can be calculated in terms of the motion vector ($-v_2^f$) and position $npos_1$ as $cpos_2 = \lfloor npos_1 + (-v_2^f) \rfloor$. Once position $cpos_2$ is calculated position $npos_2$ is easily determined using the backward motion vector v_2^b. Similar motion threading (MT) approach that employs longer wavelet filters to exploit the long-term correlation across frames along the motion trajectory was proposed in [9].

A block diagram of the 5/3 DCT temporal filter scheme is depicted in Fig.2. With 5/3 DCT filter, 1D forward 5×1 DCT transform is applied to the five connected pixels in the following order:

$$DCT\,5\times1\begin{bmatrix} x_{5t}(ppos_1), x_{5t+1}(cpos_1), x_{5t+2}(npos_1), \\ x_{5t+3}(cpos_2), x_{5t+4}(npos_2) \end{bmatrix} \quad (4)$$

From the five DCT coefficients, the three lower coefficients are named L_{3t}, L_{3t+1}, and L_{3t+2} and the two higher frequency coefficients are named H_{2t} and H_{2t+1}. These two higher frequency coefficients are considered as the high temporal frequency bands and are temporally located at the odd time positions. The temporally high-pass filtered frames can be formulated as follows:

$$\begin{aligned} H_{2t}(cpos_1) &= c_2[x_{5t}(ppos_1) - x_{5t+4}(npos_2)] \\ &\quad - c_1[x_{5t+1}(cpos_1) - x_{5t+3}(cpos_2)] \\ H_{2t+1}(cpos\,2) &= c_4[x_{5t}(ppos_1) + x_{5t+4}(npos_2)] \\ &\quad - c_3[x_{5t+1}(cpos_1) + x_{5t+3}(cpos_2)] \\ &\quad - c_5 x_{5t+2}(npos_1) \end{aligned} \quad (5)$$

where $c_1, c_2, c_3,$ and c_4 are the coefficients of the cosine matrix of order 5, namely 0.601501, 0.371748, 0.511667 and 0.195440, respectively. The lower subband frames $l_{3t}, l_{3t+1,}$ and l_{3t+2} after motion compensation are then obtained through a 1D 3×1 IDCT on the first three coefficients L_{3t}, L_{3t+1}, and L_{3t+2}, which are also down scaled by a factor of $\sqrt{5/3}$ before being transformed, as follows: $IDCT3\times1[L_{3t}, L_{3t+1,} L_{3t+2}]$. In other words, the 3×1IDCT is an orthonormal transform of

$$\begin{bmatrix} l_{3t}(ppos_1) \\ l_{3t+1}(npos_1) \\ l_{3t+2}(npos_2) \end{bmatrix} = \begin{bmatrix} 0.447214 & 0.44214 & 0.447214 \\ 0.547723 & 0.0 & -0.547723 \\ 0.316228 & -0.632456 & 0.316228 \end{bmatrix} \times \begin{bmatrix} L_{3t} \\ L_{3t+1} \\ L_{3t+2} \end{bmatrix} \quad (6)$$

The synthesis process of a sub-GOF of 5 frames is as follows: a 3×1 forward DCT is applied to the three-pixel components l_{3t}, l_{3t+1}, and l_{3t+2}, which are up scaled by a factor of $\sqrt{5/3}$ before being transformed. The resultant three frequency components L_{3t}, L_{3t+1}, and L_{3t+2} are then padded with the two reconstructed high frequency coefficients H_{2t} and H_{2t+1}. Finally, a 5×1 IDCT is

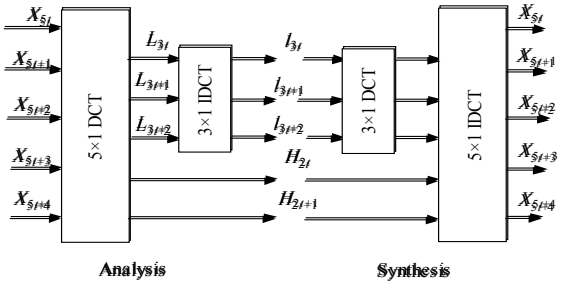

Fig. 2. The 5/3 DCT temporal analysis and synthesis.

performed on these five coefficients to reconstruct the five pixels in the sub-GOF of 5 frames. It should be mentioned that the 5/3 DCT scheme is not invertible at sub-pixel accuracy. This is because H_{2t} and H_{2t+1} may contain interpolated pixels of frames x_{5t}, x_{5t+2}, and x_{5t+4} and hence these frames cannot be reconstructed exactly. The problem of unconnected pixels in the two- and three- band structures (2/1 DCT and 3/2 DCT) was presented in [1]. We now illustrate the basic mechanism of treating connected and unconnected pixels with the 5/3 DCT temporal analysis. For all connected pixels: this case happens when a five-pixel line in each sub-GOF of 5 frames is connected, then the temporal filtering of these five connected pixels is carried out from frame x_{5t} up to frame x_{5t+4} by applying 5/3 DCT to these pixels as shown in Fig. 1. In this case, the two high and the three low subband frames are obtained according to (5) and (6) respectively. For the remaining unconnected pixels in the predicted frames x_{5t+1} and x_{5t+3}, the temporal high subbands H_{2t} and H_{2t+1} are calculated using (2). For the unconnected pixels in the reference frames x_{5t}, x_{5t+2}, and x_{5t+4}, their original values are inserted into the low temporal subbands l_{3t}, l_{3t+1}, and l_{3t+2}, respectively.

3. EXPERIMENTAL RESULTS

The performance of the longer tap 5/3 MCDCT-TF framework proposed in this paper was compared to three temporal schemes: the 3/2 MCDCT-TF, Haar wavelet, and the 5/3 wavelet filters. These various temporal schemes were then incorporated into a 3D SBC/DCT H.263+ type codec presented in [1]. This video codec consists of the MCTF (the modified MCDCT-TF or any other temporal schemes), followed by a spatial decomposition of the temporal subbands with DCT decimation technique [10]. The spatio-temporal DCT coefficients are quantized and entropy coded. In our experiments, two video sequences "*Foreman*" and "*Mobile and Calendar*" of CIF resolution (352×288 pixels, 30 frames per seconds (fps) @ YUV 4:2:0 color format) were used to assess the impact of these various temporal schemes on the coding efficiency of the 3-D SBC/DCT codec. GOFs of 8, 6, and 5 frames were chosen to generate three temporal decomposition levels with the (Haar and 5/3 lifting transform), the 3/2 MCDCT-TF, and the 5/3 MCDCT-TF schemes, respectively. These GOF lengths are so chosen that lifting under each scheme is

realized. The bi-directional block-based motion estimation was implemented (with all temporal schemes except Haar) using exhaustive-search block matching at full spatial resolution frames with a search range of ± 15 pixels. With the Haar temporal filter unidirectional motion estimation was implemented between every successive pairs of frames. The motion vectors were estimated with a half pixel accuracy at each stage of both Haar and the 5/3 wavelet filters since the lifting implementations of these filters can guarantee invertibility for any arbitrary sub-pixel accuracy. However, in order to guarantee perfect reconstruction with the 5/3 MCDCT-TF temporal decomposition scheme, motion compensation was conducted with full-pixel accuracy at the first and third temporal stages.

Fig. 3 shows the average Luminance PSNR of the decoded test sequences at various bit rates of the 3D SBC/DCT codec. The R-D curves were obtained by varying the quantization step sizes. As can be seen from this figure, the 3D SBC/DCT codec with the 5/3 MCDCT-TF temporal scheme outperforms the codec employing the 3/2 MCDCT-TF, Haar wavelet, and the 5/3 wavelet filters. For "*Mobile and Calendar*" sequence, the 5/3 MCDCT-TF scheme shows a coding gain of approximately 0.3-0.7 dB over the 3/2 MCDCT-TF scheme and this gain for "*Foreman*" is about 0.2 dB. Results of the two sequences indicate that for more predictable and uniform motion, like the one in "*Mobile and Calendar*", longer tap filter is more efficient than for sequences with non-predictable motion like "*Foreman*". It should be mentioned that in [1] the MC-3D SBC/DCT coder employing 3/2 MCDCT-TF outperforms both the H.263+ single layer and scalable coders. We should also mention that, our PSNR values are still lower than that achieved by the current state-of-the-art H.264 based scalable coding at the same bitrates. This is due to the deficiency of H.263+ incorporated into a 3D SBC/DCT codec which has lower performance than H.264. Therefore, there is a potential to incorporate these temporal schemes into the H.264 type codec to achieve a better performance than the current SVC. This motivates our future work.

4. CONCLUSIONS

In this paper, we presented a new MCDCT-TF framework for efficient and flexible temporal filtering for video coding. The proposed framework is based on the DCT transform and uses different lengths of the DCT temporal filters. The new MCDCT-TF framework was incorporated into the 3D SBC/DCT codec. The experimental results showed that longer tap DCT outperforms the shorter length DCT of our earlier MCDCT-TF as well as the Haar wavelet, and the 5/3 wavelet filters.

5. REFERENCES

[1] R. Atta and M. Ghanbari, "Spatio-temporal scalability based motion-compensated 3-D subband/DCT video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, pp. 43-55, January 2006.

[2] T. Weigand and G. Sullivan, "Draft text of final draft international standard (FDIS) of joint video specification (ITU-T Rec. H.264 —ISO/IEC 14496-10 AVC)," Joint Video Team (JVT) of ISO/IECJTC1/SC29/WG11 and ITU-T SG16/Q.6, Tech. Rep., Mar. 2003.

[3] 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.

[4] A. Secker and D. Taubman, "Highly scalable video compression with scalable motion coding," *IEEE Trans. on Image Processing*, vol. 13, no. 8, pp. 1029-1041, 2004.

[5] R. Xiong, J. Xu, F. Wu, and S. Li, "Subband adaptive motion compensated temporal filtering for scalable video coding*," in Proc. of PCS 2006*, Beijing, China 2006.

[6] A. Golwelkar and J. W. Woods "Motion-compensated temporal filtering and motion vector coding using biorthogonal filters," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 417-428, April 2007.

[7] H. Schwarz, D. Marpe, and T. Wiegand, "Scalable extension of H.264/AVC," ISO/IEC JTC1/SC29/WG11 M10569/S03, Munich, 2004.

[8] H. Schwarz, D. Marpe, and T. Wiegand, "MCTF and scalability extension of H.264/AVC," *Proc. of PCS 2004*, San Francisco, CA, USA, Dec.2004.

[9] Lin Luo, Feng Wu, Shipeng Li, Z. Xiong, and Z. Zhuang, "Advanced motion threading for 3D wavelet video coding", *Signal Processing*: *Image Communication* (*Elsevier Science*), vol. 19, no 7, pp. 601-616, 2004.

[10] K. H. Tan and M. Ghanbari, "Layered image coding using the DCT pyramid," *IEEE Trans. on Image Processing*, vol. 4, no. 4, April 1995.

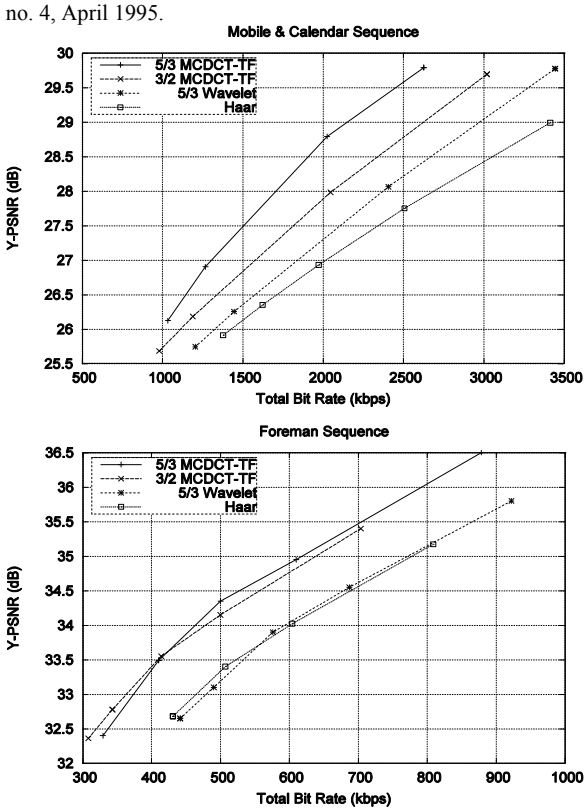

Fig. 3 Average Luminance PSNR versus bit rate of the 3D SBC/DCT codec with various MCTF schemes.